大学文科基本用书·考古文博

DAXUE WENKE JIBEN YONGSHU

Quantitative Archaeology

定量考古学

陈铁梅 编著

北京大学考古文博学院考古学系列教材之一







本书是介绍定量方法应用于考古研究的教科书，适用于考古、科技考古和文物保护等专业的学生。全书分上下两篇，上篇介绍基础统计学，下篇介绍多元统计方法。

阅读本书不要求读者有微积分知识。在讲解统计学思想、原理、方法和技术以及解读分析结果时，作者考虑到考古学生的数学基础，尽量做到深入浅出、顺序前进，并主要通过考古研究的实例、特别是中国考古研究的实例进行。这便于考古学生的理解和接受，并激发对考古资料定量思考的兴趣。学以致用是写作本书的指导原则，除通过考古实例来讲解统计技术外，作者还涉及了 SPSS 统计软件的学习和使用。

作者长期从事定量考古学的教学和研究，本书较全面地总结了十多年来我国考古学定量研究的进展，是国内第一本介绍定量考古学的参考书。

责任编辑／岳秀坤

封面设计／奇文雲海

qwyh_cn@yahoo.com.cn



ISBN 7-301-09001-3/K · 0375

定价：25.00 元

大学文科基本用书·考古文博

DAXUE WENKE JIBEN YONGSHU

Quantitative Archaeology

定量考古学

陈铁梅 编著

北京大学考古文博学院考古学系列教材之一





北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

定量考古学/陈铁梅编著.—北京:北京大学出版社,2005.9
ISBN 7-301-09001-3

Ⅰ.考… Ⅱ.陈… Ⅲ.统计学-应用-考古-高等学校-教材 Ⅳ.K85

中国版本图书馆 CIP 数据核字(2005)第 105213 号

书　　名:定量考古学
著作责任者:陈铁梅　编著
责 任 编 辑:岳秀坤
标 准 书 号:ISBN 7-301-09001-3/K·0375
出 版 发 行:北京大学出版社
地　　址:北京市海淀区成府路 205 号　100871
网　　址:http://cbs.pku.edu.cn　电子信箱:pkuwsz@yahoo.com.cn
电　　话:邮购部 62752015　发行部 62750672　编辑部 62752025
排 版 者:北京军峰公司
印 刷 者:北京宏伟双华印刷有限公司
经 销 者:新华书店
787mm×1092mm 16 开本　19 印张　423 千字
2005 年 9 月第 1 版　2005 年 9 月第 1 次印刷
定　　价:25.00 元

作者简介

陈铁梅,北京大学考古系教授,博士生导师,1959 年毕业于苏联列宁格勒大学物理系,1973—1999 年任考古系科技考古实验室主任,长期从事科技考古和定量考古的教学和研究。主要研究方向为:碳十四、不平衡铀系和电子顺磁共振测年,古陶瓷的产地溯源研究和考古资料的定量研究。发表论文近 200 篇,合作或主编专著和译著各 1 部,为建立我国的史前年代学,特别是古人类和旧石器考古年代学,为推进我国考古学研究的数量化作出贡献。曾获国家科技进步三等奖,国家教委和中国社会科学院科研成果一等奖。历任我国科技考古学会副理事长,第四纪科学研究会理事,*Quaternary Science Review-Geochronology* 和《考古科学和文物研究》等杂志编委。荣誉职称有德国国家考古研究所通讯成员等。

序　言

陈铁梅教授积二十多年从事定量考古学的研究心得和教学经验，老当益壮，以惊人的毅力写成了《定量考古学》一书。他拿着厚厚的一叠书稿给我，命我作序。我虽然不懂数学，看他的书稿也有些吃力，但仅凭一点数学常识也知道定量分析在考古学研究中的重要价值，所以很乐意在这里写几句话。

在人文科学中，考古学是应用自然科学方法和数学方法最多的一个学科。考古学是通过实物资料来研究历史的。所有实物资料都是有形和可以量度的，量的关系乃是各种事物之间十分重要的关系。通过量的关系的考察可以揭示事物的本质属性和特征，这是定量考古学得以产生和发展的客观基础。由于考古学研究的人类历史跨越数百万年，在这漫长的岁月中，反映人类社会历史的实物遗存不断积累又不断遭受自然与人为的破坏。考古学家的任务就是根据残剩下来的实物遗存来尽可能地再现已经消逝的历史。实际上这只是一个不断追求的学科的目标，要真正做到谈何容易！可是考古学家和相关的学者就是那么锲而不舍，孜孜以求，运用各种方法，包括数学方法来进行探索。残剩的实物遗存绝大多数已经掩埋在地下，需要考古学家去寻找。寻找固然要有一定的方法，更需要一个过程，一个永无止境的过程。你不可能把所有实物遗存都找到，找到的部分跟实际存在的部分是个什么关系？这里便有一个概率问题。实际存在的部分跟被长期破坏之前原本应有的部分又是什么关系？这也有一个概率问题。根据找到的遗址固然可以研究某些历史问题，但要了解得清楚一些或真实一些就必须发掘。你不可能把所有找到的遗址都发掘完，选择哪些遗址进行发掘以及发掘遗址的哪个部位，在一定程度上说是随机的。发掘的结果能在多大程度上反映遗址的整体情况，还是有一个概率问题。

在整理资料进行器物排队和分期研究时也常常遇到概率问题。比如有两种器物共存，我们说二者有同时的可能性，如果有两次、三次共存，就意味着同时的可能性比较大。如果共存的次数再多一些，意味着同时的可能性更大一些，或者用很可能、十分可能、非常可能等词语来加以说明。共存的次数达到一定数目，我们就说二者可视为同时或就是同时。这当然也是一个概率问题。我们用的词语再丰富也总是有限的，不够确切的。如果用数学逻辑来思考并用数学语言来表述就会明确得多。天气预报说今天有雨，降水概率为80%，而不说有很大可能性，就是这个道理。不过我们要明白的是，概率表述再明确也是统计性的而不是绝对的。降水概率80%自然不是降80%的水。回过来说用共存关系来判断同时性的问题。如果有三件或更多的器物共存，只要重复一两次，凭经验就可以知道它们同时的可能性非常大。共存的器物越多，需要重复的次数越少。为什么会是这样，道理很难得说清楚，可不可以用概率统计来加以说明呢！

其实考古学研究中需要运用数学的地方多得很，方法也不止概率统计一项。所有实物遗存都需要测量。大到遗址的形状大小及其与其他遗址的关系，遗址中房屋、灰坑、窖穴、墓葬、城墙、壕沟、道路等等的形状大小、分布状况及相互关系，小至一件器物的形状、

大小、厚薄和各种比例关系，人体和动物骨骼测量中的各种数据和比例关系等等，都可能并需要用数字、图表和必要的运算来加以说明。许多难以直观做出判断的事例，通过数学演算就可以有比较清晰的认识。问题在于并不是所有数量关系都可以通过初级的四则演算就解决问题的，这就需要考古学家学一点数学，学一点数量统计的知识。现在已经有一些学者试图用数学方法来研究考古学中的一些问题。例如对某些器物的类型学研究，对史家等地墓葬分期的研究，区域调查中对大量遗址及其关系的多角度研究，通过对陶瓷器或青铜器化学元素包括微量元素组成的数值变量来追溯原料产地的研究等等，都进行过一些有益的尝试。在体质人类学、动物考古学、植物考古学和地质考古学的研究中更是离不开数学方法。这些研究有的明显深化了原本的认识，有的更是开拓了新的研究领域。但也有一些研究与传统方法得出的结论不一致，甚至与常识相悖。出现这种情况可能有不同的原因，而大多数情况是对考古资料的性质认识不清，运算的前置条件设置不恰当，或者不适于用某种数学方法来处理。因此一些考古学上的问题能不能用数学方法处理，或者用何种数学方法来处理，也是考古学研究本身的问题。本书作者一再呼吁考古学家要学习和掌握基本的数学方法，正是看到了问题的症结所在而发出的肺腑之言。

本书针对大多数考古学者不甚熟悉数学方法的情况，从基本概念讲起，由浅入深地讲述考古资料定量研究的各种方法。每种方法又着重讲述基本原理、应用范围和应用方法，讲明应用这些方法的前提条件，同时说明要正确解读定量分析的结果。所有这些都结合了考古学研究中的实例，读起来不觉得枯燥和深奥难懂，反而令人有似曾相识或恍然大悟的感觉，能够引发人们运用数学方法的兴趣和自觉性。作为一部专著，本书很好地总结了我国定量考古学的进展的情况、取得的成果和存在的问题，同时介绍了国外的有关情况以供参考；内容充实，逻辑严密，图表配合也很好，在国内是第一部全面论述定量考古学的力作。

作为一部教科书，本书比较全面地讲述了定量考古学的基本原理和方法，包括使用相关软件的方法，由浅入深，循序渐进，书末还附有相关的习题，非常切合高等学校的教学和有一定基础的考古人员的自学之用。我希望本书的出版将有助于提高考古专业的定量考古教学水平，同时吸引更多的考古人员学习和掌握定量考古学的方法，促进我国定量考古学的发展，最终为提高我国考古学研究的水平而作出贡献。

严文明
2005年7月

前　言

20世纪后半叶以来，社会科学和人文科学的诸学科愈益广泛地应用定量研究和统计学方法。80年代晚期开始，北京大学和吉林大学考古系先后为高年级本科生和研究生开设了定量考古学和计算机考古的课程，这也是为了与国际考古教学的接轨。本人常年讲授定量考古学课程，甚苦于缺乏中文的教材，学生只能参考教育统计学、社会统计学等其他学科的相关教材。但是考古系学生往往是通过中国考古学研究应用定量方法的实际例子，才能较容易地理解和接受各种定量方法的原理，了解它们在考古研究中的功能和潜力。

十多年来在我国考古学的文献中陆续可见一些对考古资料进行定量研究的尝试和成果发表，如雨台山墓葬的排序，史家墓地的分期，河南早商前后陶豆的分期，侯马乔村墓地陶器分期，有胡铜戈的回归断代以及两周随葬青铜容器的组合研究等等。特别在像葫芦河流域和赤峰地区的考古区域调查中，数学方法已成为处理大容量考古资料的主要手段。这些情况反映了数量观念、概率统计观念正逐步地融入我国的考古研究中。这些进展也应该适当总结，并介绍给考古工作者。此外多种自然科学方法与考古学的结合，也必然带进自然科学所固有的定量概念和定量研究方法。例如用元素组成和同位素组成追溯陶瓷器的产地和青铜器矿源的研究，就离不开多元统计分析方法。动物考古、植物考古的资料分析中广泛应用统计学的概念与方法。考古工作者，特别是年轻的考古工作者应该对这些定量研究方法的原理有所了解并逐步应用。

编写本书的目的除作为考古系的教材外，也试图总结近年来我国考古学定量研究的进展，为考古工作者了解定量考古学提供一本参考书。本书的内容分为上下两篇，上篇介绍概率统计学基础，下篇介绍几种多元统计方法。学以致用是编写本书的原则，因此数学内容的论述尽可能结合我国考古研究的实例。作者意识到本书的读者主要是对数学不十分熟悉的考古人员，因此在编写中不刻意追求严格的数学推导，重点在于介绍各种定量方法的基本思想和原理、功能，特别是了解正确运用这些方法的前提以及对定量分析结果的正确解读。学习数学一定要实践操作，好似学游泳必须下水，因此作为附录列出少量的习题。

具有中学的代数知识和关于函数基本概念的人，应该能看懂本书的主要内容。书中在适当的章节介绍概率的基本运算法则，以及定积分基本原理等数学内容，以帮助有困难的读者。下篇的部分章节涉及初等矩阵代数，部分读者阅读会有些困难，完全可以略过不读。这些章节在目录中已用“*”号标注。

目前处理统计学的问题已有很多计算机软件，多元统计分析涉及巨大的计算工作量，必须依赖于这类软件。因此本书的第十三章简要介绍了 SPSS 软件(社会科学用统计软件包)，帮助读者入门使用。在下篇介绍聚类分析、判别分析和主成分分析的应用实例时，就是完全结合 SPSS 的有关程序进行的；重点在于帮助读者在程序执行前了解软件对

话窗口中各选项的意义和对程序输出结果的正确解读。

作者主要从事科技考古研究,虽介入定量考古学的研究和教学已20余年,但并非数学或考古学的科班出身,这两方面的学识有限,书中难免有疏漏和不妥之处。祈望同行和读者的批评指教。北京大学考古系严文明先生一直支持鼓励我从事定量考古学的研究和教学,我的年轻同事陈建立、宝文博先生阅读了全书并提出了宝贵意见,谨致谢意。

最后我引用著名考古学家、原剑桥大学考古系主任C. 伦福儒爵士对考古学研究方法的一句话作为结束语:"不计量的日子已指日可计了。"(The days of the innumerate are numbered.)

陈铁梅

2005年3月19日

目　录

上篇　考古研究中的基础统计学

下篇 多元统计方法在考古研究中的应用

上　　篇

考古研究中的基础统计学

第一章　绪　论

考古学的定量研究是指对考古现象中各种数量关系的研究,使用数学作为研究的方法。数学是研究“现实世界的空间形式和数量关系的一门学科”(辞海),属方法论学科。它既应用于自然科学,也应用于人文社会科学,包括考古科学,把数学归之于自然科学范畴是不甚妥当的。数学的最早发展起因于日常生活中的计数,随后因农业经济发展所要求的土地丈量、天文历法的研究而发展为独立的学科。我国春秋时的大政治家管仲曾说过:“不明于计数,犹如无舟楫欲径于水,险也。”但在以后的很长时期中,数学主要服务于自然科学中的天文学、物理学等所谓自然科学中的“精密科学”的研究。最近几十年来人们见证了生物科学、地质科学、心理学和社会学等长期以来以定性描述和归纳方法为主的学科研究的定量化,发展了生物统计学、地质数学、心理统计学和社会统计学等二级学科。甚至有人把学科研究中是否应用数学方法当作学科本身发展成熟程度的标志。各学科愈来愈多地注重所研究对象中量的关系的研究。这并不违背马克思关于“一种科学只有成功的运用数学时,才算达到真正完善的地步”的观点。计算机的发展和普及为数学广泛应用于人文社会科学诸学科的研究提供了现实的可能,推动了人文社会学科研究的定量化。人文社会学科的定量研究往往需要复杂的模型和大容量的计算,只可能由计算机来完成。麻省理工学院媒体实验室的创始人、尼葛洛庞帝在其著名的著作《数字化生存》中写道,“计算不再只与计算机有关,数字决定我们的生存”。另一方面,数学学科本身也是响应应用的需要而不断地发展进步的,创造出新的抽象手段,出现新的分支或方法来应付新的学科领域的要求,例如“模糊数学”以适用于很难进行精确预测的气象科学和经济科学中的课题,统计学中的“非参数假设检验”适用于处理不服从正态函数分布的变量和较低层次的有序变量、名称变量,等等。总之,数学是一门方法论学科,并不以特定的自然现象或人文社会现象的范畴作为自己研究的对象,而是服务于各个学科。

下面就(1)考古学为什么需要定量研究,以及(2)考古学研究中应用数学方法的特点和有关问题等两个方面作讨论。

1.1　考古学研究中为什么需要定量方法

考古学由于其学科特点,定量研究的开展程度远不如生物学、社会学、心理学和教育学那样普遍和深入,但也是逐年发展的。在欧美各发达国家,定量考古学或者考古统计学早已规定为考古系学生的必修课,考古学家们越来越注重所研究对象中的数量关系,定期召开关于计算机和数学考古的国际学术会议。在我国也出现了同样的趋势,北京大学和吉林大学考古系都已将定量考古学课程列入本科生和研究生的教学计划,更多的考古学家运用定量方法于自己的研究之中,在考古学术刊物中考古定量研究的论文也逐年增多。贾伟明(1987)、陈铁梅(1993),滕铭予(2000)和陈建立(2000)相继发表了关于定量

考古学研究的综述性论文。在这些论文中肯定了数学方法在考古学研究中的作用,特别强调考古学的定量研究不仅仅是使用数学方法的技术问题,而是一种思维模式,在于提倡同时注重考古学现象中质和量两方面的研究。下面我们根据考古学与数学两个学科之间内在联系,从四个方面来分析为什么考古学研究应该注重定量方法。

1.1.1 考古学研究对象内涵各种数量关系

前面已提到,数学是研究"现实世界的空间形式和数量关系",属方法论学科。考古学通过研究器物、墓葬、房址、遗址、文化类型等不同层次的遗存,或称考古实体来复原、认识古代社会,当然不可能是以单独的一件器物、一座墓葬、一个遗址作为研究对象,而是以器物的整个一个类型,以器物群、墓葬群、文化类型群作为自己的研究对象。为此要研究各层次考古实体的多种特征,包括数量特征,从中提取尽量多的信息;并要研究实体之间的各种关系、包括数量关系,由此来比较它们之间的异同,进行分类排序,这就不能排除用数学方法进行定量研究。

一种非常重要,也为考古学家所熟悉和经常应用的数量关系是百分比关系。各类型器物在器物总数中所占百分比的变化可能反映文化的地域差异或时代早晚。例如,磁山类型与裴李岗类型各遗址出土陶器的种类绝大部分是相同的。但定量分析表明,对于前者盂及支架占陶器总数的58%,而在裴李岗文化诸遗址这两种陶器较为少见,小口双耳壶和三足钵却占57%以上。这显示出两个文化类型间在炊具方面是有显著差别的。有的考古学家还以此为依据建议,它们应分别命名为两种不同的文化。

动物骨骼百分比的统计比单纯的定性研究可以更清晰、更有说服力地反映时代、生态环境的演化和社会经济形态的变革。例如,叙利亚著名的阿布胡赖拉史前遗址堆积很厚,曾分层采集了一万多片羊骨片。从下往上依次按每千片羊骨片为单位进行统计,发现在下部的继旧石器文化层中,绵羊和山羊骨片的百分比稳定在较低的6%~8%范围中,占统治地位的骨片是野羚羊,说明当时以狩猎经济为主。地层稍靠上进入前陶新石器文化早期,绵羊和山羊的骨片略有增加,在8%~16%间摆动。而在最上面的属前陶新石器文化晚期的地层中,这两种人工饲养型羊的骨片的百分比突然猛增到80%,而羚羊骨片的比重锐减。各类羊骨片百分比的统计有说服力地表明从狩猎经济向家畜饲养经济的迅速转化。分层定量统计准确地定出了发生转化的地层层位,并通过碳十四测年推断发生转化的年代(Legge etc.1986)。

考古学中的定量关系当然不局限于百分比关系。器物特征(几何尺寸、纹饰、质地)的定量研究对器物的正确分型定式也是可以作出贡献的。林沄(1980)在统计大量东北系青铜剑的基础上,总结出剑身的长宽比在总体上反映出随时代而变大的规律。欧洲青铜时代铜剑长度的统计分析观测到剑的长度参错不齐,对此可以有两种不同的解释:(1)这是因生产者和使用者不同而导致的自然涨落;(2)实际存在着长、短两种不同类型的剑。利用等长度间隔中剑的数量分布的直方图和利用方差分析方法,清楚地表明欧洲青铜时代铜剑应分成长剑和短剑两个类别。我国的考古发掘实践越来越重视地域和聚落内部遗物和遗存的空间分布,它们的空间坐标、密度和分布规律都涉及定量描述和数量研究。我国有的学者开始根据聚落中的房屋数量、文化堆积量、墓葬的人骨数和环境的负载量

等推算古人口数量，甚至进一步推算人口密度和人口增长率，探讨人口与文化进展乃至文明起源的关系。这类研究都是根据数字资料并建立一定的数学模型。考古器物与它们的原料来源的地理分布研究可以提供关于古代人群的活动范围，不同人群之间贸易交往等方面的信息。而这类研究完全离不开对所研究器物和原料的化学组成的测量，测量的结果是大量的定量数据。

一般说来，定量关系的研究比单纯的定性描述能更精确、深刻地反映客观事物的本质属性，质的抽象应以量的抽象为前提。退一步而言，定量研究至少可作为传统的定性研究的某种重要的补充。

1.1.2　考古现象与考古资料的随机性

自然界，特别是人类社会有很多现象，就其个别而言似乎是无规律的，但通过大量的试验和观察以后，其总体却呈现出明确而稳定的规律性，这些现象称为随机现象，其定量描述就是随机变量。概率统计学就是“收集和分析随机数据的科学”(《不列颠百科全书》)。

例如中国成年男性的身高就是一个随机变量。任意找一个人测量他的身高(统计学称为随机抽样)，所得结果可在 1.4 米至 2.2 米之间很宽的范围内变化，事先不能预测。但是当测量了很多人的身高后，他们的平均身高的变化范围却是很窄的。而且从大量人群的身高数据中还可以观察到一些稳定的规律，例如北方人组的平均身高比南方人组高，青年人组比老年人组高，城市组比农村组高等，反映出地区、社会进化、生活条件等因素对身高的影响。被调查的人数愈多，统计学称为样本的容量越大，反映出的规律性越稳定。不过这些规律总归是统计性的规律，并不是绝对决定论的规律，总是有一定的概率出现偏离。

考古学是利用实物遗存资料去复原古代社会的科学。考古发掘也是一个抽样过程，经常是随机的，由考古资料去推断古代社会情况，是由“样本”(局部)推断“总体”的统计推断过程，所得的结论只具有统计学的意义。举例来说在某墓地上发掘了 37 座古墓葬。表 1-1 是这 37 座墓葬按墓主人性别及有无随葬品的统计表。经计算可知男性墓葬中带随葬品的墓葬数占 69.2%，女性墓葬带随葬品的占 45.5%。现在的问题是：能否根据该墓地的资料(样本)，推断出优葬男性是该墓地所属考古学文化(总体，也是考古学研究的真正对象)的某种特征的结论？或者换一种提问方式：所观察到的男女两性墓葬有无随葬品的百分比的差别是因样品的随机性涨落引起的(例如任意抽偶数张扑克牌时，红、黑色牌的数目不一定正好相等)？抑或确实反映了该墓地所属文化优待男性的葬制？不用统计学的方法是难以正确地回答这个问题的。

表 1-1　某墓地 37 座古墓葬按墓主人性别和有无随葬品的调查统计表

	有随葬品	无随葬品
男性	18	8
女性	5	6

概率统计学用一种叫 χ^2(希腊字母，读作卡方)分布假设检验的方法来处理这一类问

题。就我们的具体例子而言,结论是:根据所观察到的该墓地男女两性墓葬有无随葬品的百分比差别不能导致“总体上墓葬有无随葬品与墓主人的性别有关”的推论。当然这个判断并非绝对正确,有一定的概率(可以计算出,约为 18 %)这个判断可能犯错误。犯错误的概率所以较大,其原因之一是因为所统计的墓葬数不够多,或者说样本的容量较小。

这个例子是有代表性的。考古发掘的资料相对于古代社会来说总是零星的资料,两者间是局部(或称样本)与全局(或称总体)的关系,因此根据考古发掘的实物遗存,推导所获得的关于古代社会的知识必然带有统计性,不是绝对真理。以这些知识作为前提进行逻辑推理所获得的新的认识同样是带统计性质的。而且古代实物遗存长期埋在地下会受到破坏,遗存的发现在很大程度上是随机的,所以英国的过程主义考古学家 D.L. Clarke(中译文,1989)在关于考古学的定义前加了修饰词,说考古学是根据“零星不完整”,而且是“被扭曲了”的实物遗存去复原古代社会的科学,用以强调考古原始资料的随机性。这不是对考古学知识体系的贬低,而是更符合客观实际。社会学、心理学等学科的知识体系同样是带统计性质的。这个例子也说明不应把定量考古学仅仅看成是应用某些数学的方法和技术,而更重要的是一种思维模式,即需要用统计学的观点,从数量的角度来看待考古学的现象和规律。

很多有见识的考古学家在自己的研究工作中确是用朴素的概率统计的观点来看待考古现象的。例如他们清楚地认识到孤证材料缺乏证明力,他们在推理论证时表现出必要的谨慎,他们在下结论时限定结论适用的范围,使用“有可能”、“有较大可能”等量词。但是我们也确实在考古文献中看到一些因不理解考古现象和考古规律的统计性质而作出的错误推论。需要指出,以朴素的概率统计观点处理考古现象总归是有局限性。概率统计学在考古学资料的定量研究中占有十分重要的地位,有时将“定量考古学”和“考古统计学”作为同一个概念来使用。因此值得提倡考古学家,特别是年轻的考古学家掌握基础概率统计的知识。或者说概率统计的观点和方法在考古学研究中是不可或缺的。

1.1.3 大信息量、复杂的考古资料需要数量分析方法

随着考古研究的深入,所累积的信息量愈来愈多,各种信息间的关系也愈益复杂。当考古学家主要凭自己的经验对有关考古遗存的肉眼观察,用自然语言对观察结果进行描述,然后对这些资料作综合分析时,难免受人脑的记忆和思维能力的限制,难以作全面的分析。通常的做法是从中找出少数几方面主要的信息,仅限于考虑少数几个,甚至于一二个变量。例如在器物的分型分式时找典型的特征,在墓葬的分期中找典型的器物。这种传统的分析程序无疑是有效的,也许是掌握了主要矛盾。但这除要求考古学家有丰富的经验、掌握大量的相关知识和花费大量的劳动外,还难免有不足之处。例如在提取主要信息的同时可能把大量被认为是非主要的信息丢失,而且不同的考古学家对同一问题作研究时所选取的典型特征、典型器物可能是不一样的,研究的结论也会有所差异。这种情况下难以严格地判断哪一种结论更符合实际。在某些研究中,所谓的典型特征或典型器物似乎是在对研究结论已有某种先验的看法的情况下“选取”的,因果关系有可能被颠倒。

多元统计方法,或称多变量分析方法能方便地处理大批量的而且复杂的考古信息资料,包含大量的考古实体,每个实体具有多方面的特征(变量)。在处理这类复杂问题时,多元统计方法可对实体排序、分类,还可以揭示变量与变量之间的相关关系。处理有关数据时可以给某些变量加不同的权重。加权实际上颇类似于传统方法中的选取典型特征,但选取的标准不单是凭研究者的个人经验和主观认识,也可以由数据结构本身导出。多元统计方法在处理数据时,除给"典型特征"加权外,也同时考虑所有其他的特征,减少了顾此失彼和绝对化的缺点。用多元分析方法处理数据都是借助于计算机的,数据处理计算过程甚快,适当地改变重点特征的选取和改变所加权量,很快又能得到另一个分析计算结果,供研究者分析选择。

多元统计方法的应用和其他定量方法一样首先要求对被研究实体的诸特征作定量化描述,由于不同类型考古实体其特征的定量化描述的难易程度不同,目前多元统计方法在考古学不同类型的研究课题中取得的进展和成功程度也很不同,这在本章下面 1.2 节中将详述。

在区域考古调查中需要综合处理在不同地点、不同自然和人文环境下出土的不同时代的不同遗存的大量信息,需要研究各个变量之间的关系,这是离不开数量分析方法的,正如在葫芦河流域和赤峰地区考古调查的考古学家们所做的那样。随着考古资料的增多,人们正在建立各地区的考古学地理信息系统,这同样需要将已掌握的大量考古资料定量化符号化,并按规定的原则输入地理信息系统。

1.1.4 数学是一种特殊的语言系统,是自然语言的补充

数学也是一种语言、符号系统,它经常应用图、表等工具表述现象和规律,这种表述方式具有简明清晰等优点。例如本章前面的表 1-1 是一张 2 行 2 列的表,它等效于自然语言中的四句话,即按行横读 2 句:"男(女)性墓葬中带随葬品的有 18(5)座,无随葬品的有 8(6)座";按列竖读 2 句:"带(不带)随葬品的墓葬中墓主人为男性的有 18(8)座,墓主人为女性有 5(6)座。表 1-1 的行数和列数均不大,表格语言简明的优点也许还不太明显。我们设想有一张记录 25 座墓葬中(行)15 种器物数量分布(列)的统计表。该表就相当于 25 加 15 共 40 句陈述句,反映每座墓葬中含有哪几种器物,每种器物多少件(横读 25 句);和每种器物出现在哪几座墓葬中,数量多少(竖读 15 句)。表格比自然语言简明多了。而且,如果这 25 座墓葬在表中已按序分期排列,从表中还可清楚看到每种器物按时间演化的规律,并可反过来按表中器物演化的规律来判断墓葬的分期排列是否正确。这在第十七章关于史家墓地 14 种器物式别在 6 种分期方案中的分布比较表中可以清楚看到(见表 17-8 和表 17-10)。

图形也是数学语言的一种重要表述方法。图 1-1 和图 1-2 分别表述了青海柳湾墓地各死亡年龄段人骨的百分比,男女分别统计。图 1-1 用的直方图,而图 1-2 用累计百分数曲线。前者清楚地显示了青年女性的高死亡率这个特点,而后者清晰地表明男性的平均寿命高于女性。

前述图表表明,图表语言比自然语言不仅简明清晰,而且更容易揭示出所观测数据中所隐藏的规律性。

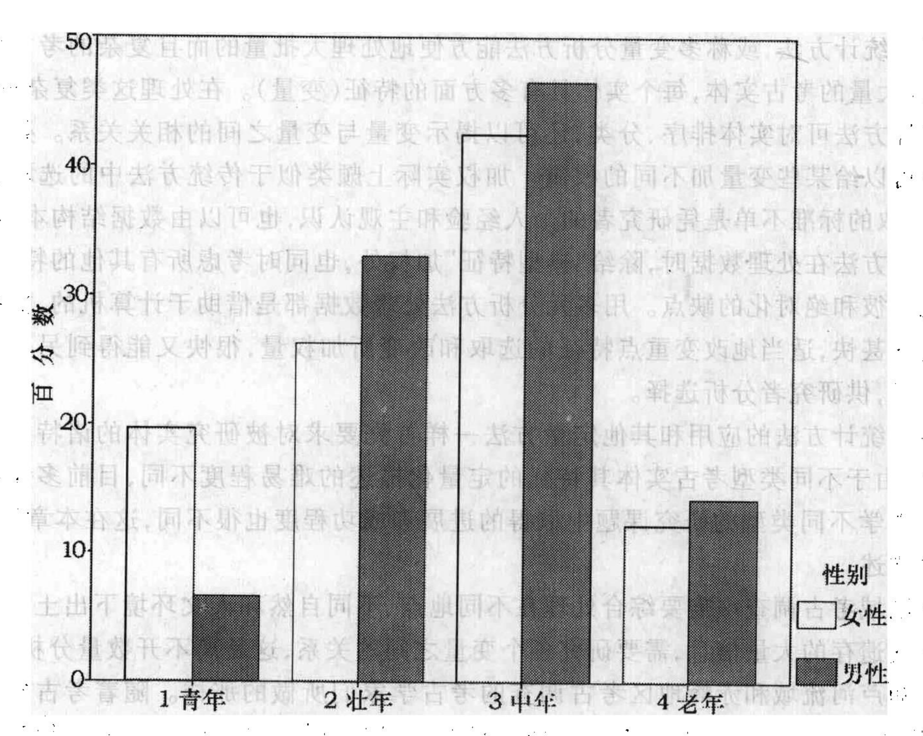


图 1-1 柳湾墓地男女人骨按年龄段百分数分布的直方图

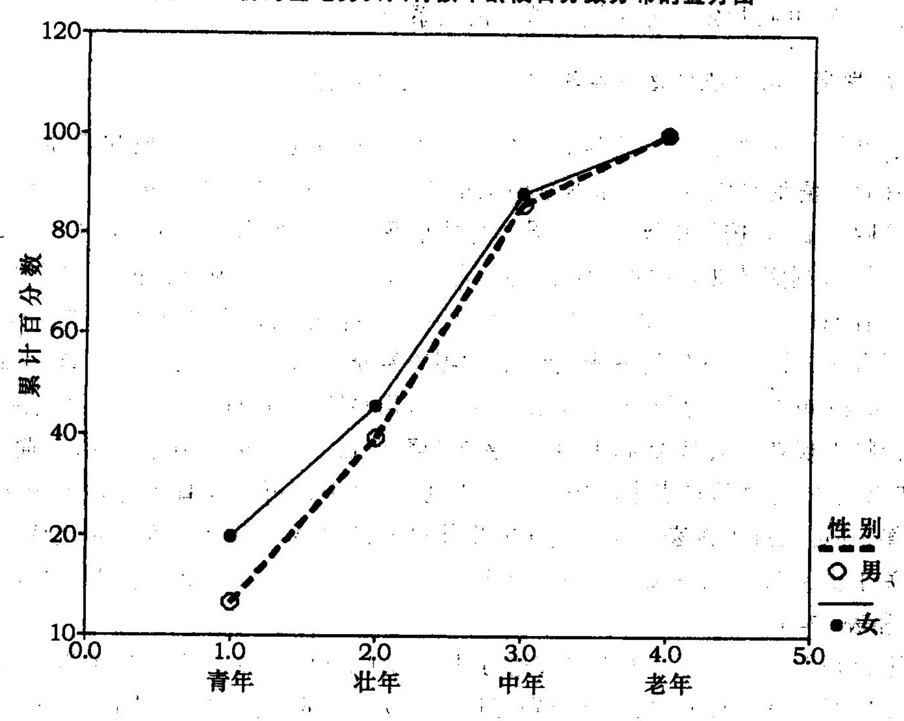


图 1-2 柳湾墓地男女人骨按死亡年龄段的累加百分曲线

我们建议考古工作者多使用数学语言，还出于以下两方面考虑。(1)各种自然科学方法，如测年、文物的成分分析、生物学技术和遥感等，在考古研究中的应用愈益普及，更多的自然科学工作者已成为或正在成为考古学家的合作者。自然科学工作者在自己的

研究工作中,在发表研究成果时,是离不开图表等数学语言的。考古学家熟悉数学语言能促进与自然科学工作者的交流合作。(2)计算机在处理,贮存考古资料中的作用已为大家所承认,应用也日益广泛,但计算机主要处理数字和符号,目前它处理自然语言,特别是汉字的能力还有限,很多软件都是为处理数字和符号编写的。

1.2 考古学研究中应用数学方法的特点和有关问题

前面我们从考古现象中的定量关系、概率统计的观点、复杂资料系统和辅助的语言符号系统等四个方面论述了数学在考古学研究中所能起的作用。本节将论述数学方法应用于考古学的特点和我们在这个问题上的某些看法。

1.2.1 定量研究作为一种思维模式要求考古学家的亲身实践

数学并不是以某特定范畴的自然现象或社会现象作为自己的研究对象,而是研究自然和社会现象中各种数量关系的一种抽象体系,属方法论学科。因此它在考古学研究中的应用,不同于碳十四测年、孢粉分析或文物的化学成分分析等。它不是一门分离、独立的专业技术,而是考古学家为整理考古资料,或者更确切地说,是整理自己头脑中对考古资料的认识的一种工具和思维模式,是每天每时都要使用的工具。考古学家不必亲自去观察孢粉,也不必亲手去测定碳十四样品年龄,所要求考古学家的主要是正确采样和合理地分析测试结果(这又要求用数学)。但是考古学家必须自始至终并亲自清楚地意识到他所研究课题中的数量关系,他所掌握资料的概率统计性质。他应有能力从自己的课题中提出数学问题,在与数学家和统计学家合作时能向他们解释清楚所涉及问题的考古学内涵,与他们商讨采用合适的数学方法处理考古资料,共同认识隐含在考古资料中的规律、模式。总之,考古学家必须亲身参与考古学资料中定量关系的研究,别人是不能代庖越俎的。那种认为考古学家提供资料,请数学家来做定量分析的看法是完全错误的,多数情况下不可能得到什么有意义的成果,往往是数学家不切题的答案对应于考古学家模糊不清的问题。与数学家的交流合作要求考古学家本身应具有一定的数学知识。

笔者认为未来对定量考古学作出贡献的将是考古学家本身,因为只有他们才懂得所研究考古现象的真正内涵,懂得数学计算结果的考古学含义,懂得选择、甚至发展合适的数学工具去解决考古研究中的特殊问题。已走在前面的西方考古学界的现实情况也是如此。笔者寄深切希望于我国年轻的考古工作者,只有他们才能真正推进我国考古学的定量研究。

1.2.2 定量思维贯彻于考古研究的各个阶段

定量方法并非只是在考古研究的最后阶段,即资料整理和总结阶段才被运用。在制定研究计划时,研究者不仅决定要做什么,还要决定怎样做,发掘阶段收集什么资料,怎样收集资料等。这些先期的计划应与事后用怎样的方法来分析资料是相配合的。因此定量方法、或更确切地说定量思维的模式,在考古研究的整个过程中应得到自始至终的贯彻。例如在本章 1.1.1 节介绍阿布胡赖拉遗址羊骨片统计的例子中,为了最后分析人

工饲养型羊骨片百分比随时间的变化,在发掘时就必需仔细分层发掘,而且清楚标明每片骨片出土的层位及埋藏深度。如果到最后资料整理阶段才发现资料不全或不够详细以至影响进行深入的定量研究,往往已为时过晚,难以弥补了。一般说来如果考古工作者自始至终自觉地注意定量关系,会使得他的发掘工作、发掘记录更仔细、更科学。

1.2.3 定量研究方法并不难,可以借助计算机的帮助

在运用多元分析方法处理复杂繁琐的考古资料时,数学处理过程看起来似乎很复杂,甚至难懂和吓人。其实在很多情况下,即使没有数学家的帮助,考古学家自己用某些简单的数学手段就能得到有意义的分析结果。例如1.1.4节中用图表方法对柳湾墓地各死亡年龄段男女人骨百分比比较的表述。考古学家黄蕴平(1996)曾自己计算了周口店第一地点和南京汤山两地出土的肿骨鹿腿骨直径的平均值和标准差并作统计检验,发现前者粗壮而后者纤细,由此得出两地的相应地层在时代上可能有先后的考古推论。考古学家袁靖等(Yuan,2002)根据山东一些贝丘遗址各层贝壳的平均尺寸及其标准差做统计检验,由此推论因生存压力导致晚期贝壳平均尺寸的变小。这里计算平均值和标准差是简单的算术运算,做统计检验也不复杂,考古学家不难自己动手实现,何况可以使用现有的计算机软件。即使复杂的多元统计分析,也有相应的软件,完全不需手工计算。关键是要定性了解计算方法的基本原理,学习正确地使用相关的计算机程序和对程序输出结果的正确解读。可喜的是我国部分中青年考古学家已经在自己动手操作。

1.2.4 考古学定量研究的初期阶段犯有错误是难免的

考古学学科在刚开始应用定量方法,特别是各种统计方法时,难免犯各种各样的错误,有时甚至是很原始、低级的错误。这并不可怕,甚至有一定的必然性。最常见的错误是使用某种统计学方法或程序时,忽略了每种方法或统计软件在处理数据时所要求的前提条件,另外就是所处理的数据样本容量太小,等等。这类错误在我国近几年发表的考古论文中也是能看到的,例如有的论文对某文化类型仅测量了一片陶片的化学组成,就把它看成是该文化类型陶片化学组成的代表作聚类分析,其研究结论的可信度肯定成问题的。还有在用联列表的 χ^2 检验来判断两个考古学因素之间是否有关联时,联列表中部分单元格中的个体数太少导致了判断结果不稳定。这些情况说明要求普及定量考古学的基本知识,以便准确地应用定量方法。

美国于20世纪60、70年代,过程主义考古学盛行。过程主义学派注重考古现象之间的关系研究和提倡假设检验方法,这必然促使他们应用概率统计学中的各种关联研究和假设检验的理论与方法,他们还把考古研究的定量化与客观化,科学化相联系。因此当时很多考古学家在自己的研究中积极运用数学方法,但在运用中也犯了很多错误,甚至出现个别例子因脱离考古学内容乱用数学方法而闹笑话。80年代初后过程主义考古学兴起,该学派在批评过程主义考古学的错误时,曾一度对过程主义考古学家常用的定量方法本身提出怀疑。但在1985年美国考古学会第50届年会上,组织了关于定量考古学的专题讨论,总结这方面的经验教训。会后由美国西北大学人类学系的M.S.Aldenderfer(1987)教授主编出版了该专题讨论会的论文集。这是一本定量考古学方面很有影响的

书。90年代以后美国的考古界对考古学的定量研究逐步建立一种积极而正确的态度。

1.2.5 定量研究不排除主观性,它与传统的考古研究方法是相辅相成的

数量关系的研究不仅能揭示被研究考古资料中内含的,而不易被传统的定性研究所看出的某些现象和规律,而且定量研究排除了在归纳和演绎等推理过程中可能出现的主观任意性。定性研究的过程使用自然语言,在推理过程中有时所涉及的概念的界定不一定严格,在表述者和读(听)者之间不一定有完全的共识,此外,推理过程中逻辑关系的表述也不一定清晰和连贯。定量研究的特点是概念界定清楚,逻辑严格。但定量方法并不是笼统地排除主观性,在对材料(数据)的选取,对材料中不同组成部分所给权重的大小,采用哪一种定量方法,处理材料时考虑哪些因素,哪些因素不予考虑等方面,研究者都可以凭自己的知识与经验自由选择和决定。这是研究者的主观意见。但材料与方法一旦选定后,下一步对材料(数据)的处理过程就必然要按照严格的数学逻辑来进行,中间不可能任意变动。即数据和处理方法确定后,最终的结果也已确定了。因此在定量研究中研究者的主观意见是放在桌面上的,大家看的见,明晰的而不是隐藏的。数学方法排除的是主观任意性。

另一方面也需要指出,传统的考古研究方法善于在大量的考古材料中,充分利用考古学家对于有关研究课题已有的知识和经验,通过反复的对比分析,排除干扰,寻找典型特征、典型器物,即掌握主要矛盾。在研究像陶鬲等形态复杂的器物时,掌握主要特征是十分重要的。此外它还考虑考古现象的特殊规律,如早期的遗物可以在晚期的单位中出现,类型学中的祖型和遗型问题等。目前的各种定量研究方法虽有前述的各种优点,但还不能理想地处理这类问题。此外对器物外型的定量描述有时比较困难,例如对鬲这类非辐射对称形的器物。鬲的裆高虽可以直接测量,但是应该使用裆高的直接测量值,还是用裆高与鬲总高的比值更合适呢。另外怎样对鬲足形状的定量描述等都比较困难。总之定量方法本身还不完善,需要改进发展,考古研究中传统方法和定量方法是互相补充的。在第十七章我们将通过详细比较两种方法对史家墓地分期的异同,进一步阐明它们的互补性。

1.2.6 定量考古学的教学是与国际接轨、与自然科学工作者合作的需要

为了在考古学研究中注重和开展定量关系的研究,考古工作者,特别是青年考古工作者和大学考古系的学生应该学一点数学。美国、西欧及前苏联各主要大学的考古系和人类学系都开设定量考古学(或考古统计学)等课程,出版了多本定量考古学的教科书。例如英国伦敦大学 Shennan 编写的《Quantifying Archaeology》1988 年出版后,1990 年再版,而 1997 年又出版了第二版修订本,牛津大学考古学院 Fletcher 等编写的《Digging Numbers》于 1991 出版后,也几次修改再版。与我国有密切合作关系的匹茨堡大学人类学系主任 Drennan(中文名为周南)亲自讲授定量考古学课程,并编写教材《Statistics for Achaeologists》(1996)。前苏联高等教育部早在 1987 年就审批出版了高等学校定量考古学教科书(Федоров-Давыдов, 1987)。考古工作者要学一点自然科学知识和有关的数学知识早已成为国外考古学界的共识。近年来我国考古学家投寄国外考古学术刊物的论文,常被编辑

和审稿人要求对数据作规范的统计学处理和表述。英国剑桥大学考古系前主任、著名考古学家 C. 伦福儒爵士十分重视考古资料的定量研究,他认为在考古研究中“不计量的日子已指日可计了”。90 年代初作者访问伦福儒时,他曾谈到中国学生如到剑桥读考古学位,要补的第一门课程是概率统计学。

在高等教育中,西方国家的经验是应该单独为考古系学生开设定量方法和统计学的课程。虽然用于各学科的定量方法和统计方法很多是共通的,但考古系学生在其他系旁听这方面课程的效果并不好。因为这些课程中所举的实例与考古学毫无关系,考古系学生感到这些例子陌生,不好理解,甚至使人厌烦。很多考古系学生的数学基础不强,数学本身已够难学了,不应再附加这些陌生的例子。已有的教学经验表明,不熟悉数学的人往往更易于通过自己所熟悉的实例来逐步把握定量方法的内容,而不是首先掌握作为应用基础的抽象数学逻辑。成功应用于考古研究的数学方法不仅易被考古系学生掌握,而且能引起他们学习数学方法的兴趣,激发联想,产生在自己的研究课题中关注定量关系的愿望。

近年来国家教育部已审定以初等微积分为主要内容的高等数学列为高等院校文科一年级学生的必修课程,反映了当前学科综合和文理结合的总趋势。北京大学和吉林大学考古系先后于 80 年代末和 90 年代开始讲授定量考古和计算机考古课程,有部分同学因为认真学习而受益。另一方面约 20 年来在我国考古学的研究中,定量研究方法经历了虽然艰难、但不断进展的探索历程。我们欣喜地看到一些中青年考古学家已关注自己的研究课题中的数量关系,并正确、有效地用数学方法来研究考古资料中的数量关系。据不完全的统计有葫芦河流域和赤峰地区考古调查,某些旧石器地点石器和石片的分布规律研究,雨台山墓葬的排序,史家墓地的概率方法分期,有胡铜戈的回归断代,乔村墓地出土器物的聚类分析以及两周随葬青铜容器的组合研究等,这些工作都是由考古学家亲自完成的。此外在我国有不少考古学家从事动物考古、植物考古和体质人类学的研究,他们的研究资料原本就包含了描述被研究对象形态特征的长度、厚度、角度、比值指数等数值型参数,各类研究对象出现的频次、频率等定量数据。因此这些考古学家已普遍使用概率统计学的基本方法,而且在使用传统的对多个单项指标进行统计比较的同时,逐步应用聚类、主成分分析、相关分析等综合的多元统计方法。体质人类学中应用多元统计方法的优点在对山东大汶口、安阳殷墟、辽宁喇嘛洞墓地、新疆营盘墓地颅骨的种族关系研究等工作中得到了充分的体现。至于目前主要由自然科学工作者进行的关于古陶瓷产地的溯源研究,因研究是基于测量古陶瓷化学元素组成的数值变量,近年来发表的几十篇论文全部使用多元统计方法,除常用的聚类、主成分分析、判别分析等方法外,还尝试使用了人工神经网络、模糊聚类等方法。文理学科的结合,考古学家与自然科学学科的人员愈益紧密的合作也紧迫地要求中青年考古工作者和考古系的学生学习定量考古学的内容。

本书作为我国第一本关于定量考古学的专门著作,作者希望它能对十多年来我国考古学研究中应用定量方法的进展、成果和问题有所总结。同时作为一本教材,应能帮助学生掌握考古定量研究的各种基本方法。本书不是一本讲述基础概率统计学和多元统计分析的数学书,因此不拘泥于严格的数学推导,经常不通过推导、证明而直接给出最终

的数学公式。但本书更不是“烹调指南”式的工具书。本书将系统地、由浅入深地介绍应用于考古资料定量研究的各种方法的基本思想和原理、功能，特别是阐明正确运用这些方法的前提条件和对定量分析结果的正确解读。全部的内容论述都将结合实际的例子，特别是我国考古研究中的实际例子。在很多章节还介绍如何应用 EXCEL 和 SPSS 等计算机软件于实际问题的解决。

第二章　考古资料的定量描述

2.1　考古实体和实体的属性

考古学研究器物、墓葬、房址、聚落、遗址和文化类型等不同层次的文化遗存。为了对各类文化遗存作定量研究，首先要对遗存资料作定量化的描述。本书中称这些遗存为不同层次的考古实体，实体有时也称为样品、个体或个案等。我们知道。考古学不是研究单件的器物、单一的墓葬，而是将器物的整个一个或几个类型、一组或几组墓葬群，一类或几类文化类型作为自己的研究对象。为此要研究和比较同一层次不同实体之间性状特征的异同，根据性状的异同来对实体群进行分类和排序。譬如说一批陶豆可以根据其形状、纹饰和陶质来分类。形状，纹饰和陶质等描述陶豆性状特征的各个信息项目称之为陶豆的属性。对考古实体的定量研究，首先要对其属性作数量化或符号化的描述。数量化或符号化描述的属性称之为观测数据。因为同一属性在不同的实体上反映为相同或不同的观测数据，因此观测数据也称变量。变量和数据在本书中经常作为相同的概念使用。描述实体群的有关属性的观测数据或变量组成了考古学定量研究的基础资料。

需要指出，实体和属性是相对的概念。例如把某种器物究竟看成是实体还是属性，要依据所研究的问题而定。如果希望对这类器物进行分类，该类器物的各个个体就是实体；如果这些器物是随葬品，并作为墓葬群分期的依据，相应的器物就是其出土墓葬的属性了。

2.2　属性的定量描述和数据的类型

考古实体的属性是多种多样的。描述实体属性的数据类型基本上可以分成三类：名称属性或名称变量，有序属性或有序变量，以及数值属性或数值变量。以陶豆为例，其纹饰和陶质等属性是定性属性，很难用数值来描述；而反映其形状的诸属性，如陶豆的高度、其盘直径、盘深、柄、底高、底直径等可以用数值来描述，属数量特征。因此考古资料定量化以后的数据类型是不一样的，处理不同类型数据的数学方法也是不一样的。

2.2.1　名称属性或名称变量

实体的某些属性反映为若干种不同的状态，而对于每个个体，只能是处于其中的一种状态。例如陶器的底部形态可以是平底、尖底、球面底、圈足底等不同的状态。其纹饰可以有绳纹、蓖纹、玄纹、指刻纹等。但每个陶器实际的底部形态和纹饰只能是固定的一种，例如平底和蓖纹。描述陶器底部形态和纹饰的属性就是名称属性，名称属性也称为名称数据或名称变量。对陶器的底部形态这个名称属性的定量化，可以用 1、2、3、4 四个

数字依次代表平底、尖底、球面底和圈足底，即对于某尖底的陶器，描述其底部形状的变量的取值是2。这里这些数字仅仅是符号，它们之间不存在大小的关系，也不能进行一般的算术加减运算，数学运算符“>”和“<”也不能应用于名称变量。我们同样可以用A、B、C、D四个符号依次来描述平底、尖底、球面底和圈足底。因此名称变量的定量化实际上是符号化，而不是数值化。定量化是比数值化更为广泛的一个概念。

需要指出，对于所研究的实体群，描述其某个名称属性的符号的数目应组成一个完整的集合，同时各符号所描述的状态间应该属同一层次并是互斥的，即每个实体必须处于其中的某一状态，同时也只能处于其中的一种状态。例如，我们不能用有纹、无纹、绳纹、蓖纹四个状态组成一个“组合”来描述陶器的纹饰。因为有纹与无纹已组成完整的集合，它们又是互斥的。至于是绳纹或蓖纹，那是有纹状态的下面一个层次的属性。但是如果把无纹即素面看成纹饰的一种状态，那么素面、绳纹、蓖纹和方格纹等可组成同层次属性的完整集合。名称属性是考古学中常常遇到的。属性的定义和应包括哪几种状态才组成完整的集合，应是考古学家根据具体的研究内容来确定的。但定义必须明确清晰，不能含糊不清。

名称变量中的一种特殊情况是二元变量。这里属性只能处于两种状态之中的一种。例如人的性别必须是男或是女。又如某动物种在某个动物群中出现或缺失，器物有无纹饰等，两者必居其一。这两种互斥的状态组成了完整的集合。二元属性的取值一般用“0”和“1”两个值来表示。二元属性是考古学中常见的属性，本书后面将介绍一些专门的数学方法来处理二元变量。

2.2.2　有序属性或有序变量

有序属性与名称属性相似，也反映为实体的某个属性可以处于多个不同状态中的某一个，其不同之处是有序属性的各状态之间有一定的顺序关系。例如在分析墓葬出土的人骨的年龄组成时，把人骨按年龄段分成婴儿、儿童、少年、青年、壮年、中年和老年等7个的状态，这7个状态是有顺序关系的，因此人骨的年龄段是人骨的某种有序属性。这里也可以用1、2、3、4、5、6、7等七个数字来相应地描述这七个状态，这时数字的大小反映顺序的位子，数字之间的差反映两个状态之间相隔多少位。它们之间的减法运算也是有意义的，表示2个状态相距几个序位。数学运算符“>”和“<”对有序变量也是有意义的，它们表示状态的先后。但是加法和乘除法运算对于有序变量却是没有意义的。

考古学研究中常见的有序变量有器物和墓葬等实体的分期、地层的次序（第几自然层或第几文化层）和沉积物的粒度（黏土、粉沙、细沙、粗沙、砾石）等。

应当指出，有序变量两个相邻状态之间的“差距”只表示顺序的次序，而不表示在时间或空间上的数量差距。举例来说地层有一定的次序，但各层的厚度可以不一样，其堆积延续的时间跨度也完全可以不一样。

2.2.3　数值属性或数值变量

可以用数值来描述的属性称为数值属性或数值变量，这是最常见的一种属性。各墓葬出土的某种器物的数量及其百分比是数值属性，器物的几何尺寸和重量，陶瓷和青铜

器化学组成中各元素的含量都是数值属性。数值变量可以直接参与通常的各种数学运算。数值变量又分成离散型和连续型两类，如某种器物出现的次数只能是1、2、……等正整数表示，属离散型数值变量；而物体的长度、重量等则可以用小数来表示，就是连续型数值变量了。有的书中还把数值变量分为比例量和区间量，在本书中不拟详细讨论它们之间的差别，数值变量中的绝大多数是比例量。

2.2.4 变量的层次和数据类型之间的转换

上面所述的三种变量的层次是不同的，从高到低依次为数值变量、有序变量和名称变量。为什么要注意属性或数据的类型呢？这一方面反映了实体的属性本身的性质不同，另一方面各种数据处理的方法对数据的类型有一定的要求，处理数值数据的方法不一定能用于处理名称数据，反过来也是这样。例如数值变量间的关系用皮尔逊相关系数表征，有序变量间用斯皮尔曼等级相关系数，而 Φ 和 V 关联系数则表征名称变量间的关联强度。需要说明变量层次的高低与变量在课题研究中重要性的高低并没有必然的联系，层次偏低的名称变量在考古学研究中经常起到十分重要的作用。

不同类型的数据在某些情况下是可以转换的。例如沉积物的粒度是有序变量，但如果我们用黏土、粉沙、细沙、粗沙、砾石等粒度直径的平均值来描述，就是数值变量了。一群人的身高是数值变量，但把身高以每10厘米分段来描述，就是有序变量了。多状态的名称变量如颜色：红、黄、蓝、绿、白，有时可用红与非红两种状态来描述，这时多状态名称变量就转化为二元变量了。二元变量的一个优点是，经过数据标准化后，它可以与数值变量一起参加运算。（关于数据的标准化本书后面会介绍。）

2.3 考古器物形状的定量描述

本节将介绍对器物形状作定量描述的几种方法。陶器和青铜器是考古学研究中最常见和重要的实体。器物可分成瓶、鬲、罐、豆等器物种类，是依靠考古学的直观知识和传统进行的，一般容易得到共识。器物的名称如瓶、鬲、罐、豆等也就成为判别器物分类属性的名称变量。每种器物的分型定式是判断考古学文化的地区类型和时代的重要因素。但器物准确的分型定式却远非简单的任务，这依赖于考古学家的经验和学识。能否对考古器物的形制作定量描述，并在此基础上对器物分型分式呢？这是比较困难的。考古器物的形制属性一般是名称数据，如口沿的形式、颈、腹以及底部的形状等，较难用数值来表示。但有些轮制的器物因辐射对称，有较为简单有效的方法对其形状作数值描述。下面举例说明对轮制器物的数值化描述。

例一，轮制器物的定量描述方法之一。图2-1所示为一轮制的似圆柱形瓶，可以把它的高度分成6等份，然后记下高度h和高度6等分处7个截面的直径数值 d_1，d_2，d_3，d_4，d_5，d_6，d_7。其中 d_1 是瓶的口径，d_7 是底径。高度加上7个按次序排列的直径值，这8个按次序排列的数值变量就组成描述瓶的形状的一组数据，写成：$(h, d_1, d_2, d_3, d_4, d_5, d_6, d_7)$。对每一个瓶而言，这8个数或这一数组是唯一确定的；而不同的瓶用不相同的另外8个数来描述。每个瓶与每组数据间是一一对应的关系。如果需要对瓶的形状描述

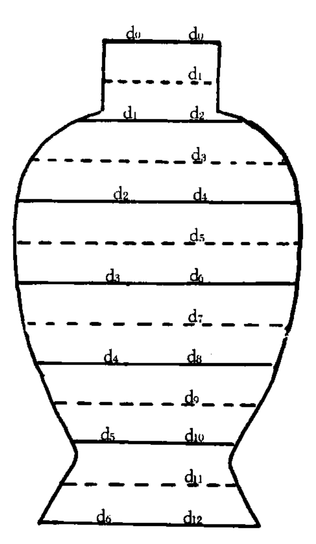


图 2-1　对辐射对称形器物形态的一种数值描述方法

更精确，可以增加沿高度方向的等分点，譬如说分成 8 段，10 段或更多的段。这时数组中变量的数目也相应增加，使得以后进行数据处理时花费的时间和经费也要增加。这里需要研究者的决策，分成几个等分较妥，取决于瓶子形状的复杂程度。像图 2-1 所示的瓶子分成 6 段到 8 段就够了，基本上能反映出瓶口处是直的，其腹部最粗处的大致高度和直径值，瓶子最细处的大致高度和直径值等。但这种描述方法也有一定的缺点。考虑两个形状相似，只是大小有差别的瓶子。上面这 8 个变量的取值对这两个瓶子都会有明显的差异，不能合适反映两瓶子间形状的相似性。为了解决这个问题，可以用各直径和高的比值(d_i/ h)来替代直径本身的数值。这时描述瓶的大小和形状的 8 个数值为(h, d_1/h, d_2/h, d_3/h, d_4/h, d_5/h, d_6/h, d_7/h)，对于两个形状相同仅大小不同的瓶子，8 个变量中仅第一个变量 h 不同，它反映瓶的高矮不同，而后面 7 个描述瓶子形状的变量是相同的。

例二，轮制器物定量描述方法之二。对轮制器物形状的定量描述经常可根据其外观特点和考古工作者的常识找出一些形态特征来描述。例如本书作者(陈铁梅等 1989)于 80 年代中期在对中原地区从二里头期到人民公园期的 13 件陶豆进行分析时，曾用了下面 6 个变量(见图 2-2)对陶豆形状作定量描述。它们是:(1)陶豆的高度——通高，(2)口径与通高的比值，(3)柄高与通高的比值，(4)盘深与通高的比值，(5)陶豆最大直径与最小直径的比值，上面 5 个是数值变量；第 6 个变量描述陶豆有没有纹饰，用二元变量的两个取值 0 与 1 表示。然后用主成分分析方法处理了这样定量化后的数据，13 件陶豆被分成三组，基本上与二里头期，二里岗早、晚期相对应，只有一件器物是例外。在上面陶豆形状的定量化过程中，不仅应用了通高，直径等直接度量陶豆大小尺寸的线性尺度，而且应用了一系列线性尺度量的比值，这些比值反映了对陶豆形状一般常识，例如口径与通高比反映陶豆的胖瘦，柄高与通高比反映是高柄豆还是矮柄豆，最大直径与最小直径比反映是粗柄豆还是细柄豆。选择考古器物的常识性特征是定量描述器物形状常用和很有效的方法。滕铭予(2004)对侯马乔村墓地陶器分期时也采用了类似的方法来描述器物的形状。

例三，对于非圆形辐射对称的器物很难用上述的办法来定量描述。在有的定量考古学的书中介绍一种称为马赛克的方法。大致是把器物的外型轮廓画在一张方格纸上，每个方格是编号的，以轮廓线是否通过某个方格用 0、1 表示。纸上有多少个方格就用多少个 0、1 数来描述这个器物。如方格小，描述更精细、确切，但方格的总数就多，数据量变大。本书不认为马赛克方法很适用于考古器物的描述，因为数据量太大，计算起来十分复杂不方便，更重要的是不少器物的外型轮廓是与观察的角度有关的，从不同的角度观测，器物的外型轮廓线是不一样的。再之，马赛克方法没有充分考虑考古器物常识性的特征。总之对非圆形辐射对称的器物，特别对陶鬲等形状复杂的器物准确有效的定量描述是一个需要进一步研究的问题。

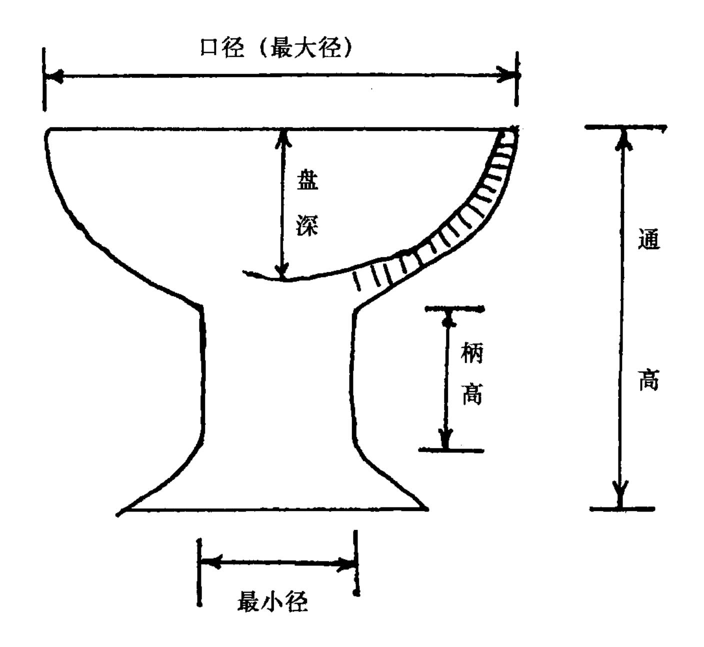


图 2-2 对陶豆形态数值描述的示意图

2.4 考古实体的描述中属性的选择

每一类考古实体的属性是多种多样的，不可能也没有必要对其所有的属性都进行描述和研究。属性的选择决定于所要研究的课题。例如发掘了一个墓群，要研究的问题是该氏族墓地贫富分化的情况。这样墓穴的大小，每个墓出土的随葬品的数量和质量等是最需要注意的属性，此外也许还要注意墓主人的性别。墓的位置和朝向可能不是最重要的属性。如果要研究墓地中各墓葬的时代早晚，那么墓葬在墓地中的位置，随葬器物的式别就成为应该重视的属性了。因此，属性的选择是需要考古学家根据其研究目的和知识来决定，没有人可以代庖越俎的。对选定的属性怎样定量化同样需要由考古学家的知识来决定的。

这里要强调两点：(1)我们不是对单个实体进行定量化描述，而是对整个一类实体作定量化描述，是对整个一类实体的某个或某几个特征的多种表现形态作定量描述。(2)不同的考古学家对同类实体作定量描述的选择可能是不同的，因为他们可能有不同的侧重点，选几个属性，怎样对属性进行定量化，都由考古学家自己来决定，依赖于考古学家的知识和个人的选择。因此定量描述的方法不是唯一的，不排除研究者的主观倾向性，不过这种主观性，不是随意和隐藏的，而是摆在桌面上的，大家都看得见的。

2.5 原始数据统计表和计算机电子表格软件

对一批同层次的考古实体的有关属性经选定，并加以定量描述后，所得的数据用表格的形式表示出来，就是原始数据统计表。这是完整的，定量化的原始资料，是以后对这

批考古实体进行研究分析，如分类，排序，比对的基础资料。表 2-1 是原始数据统计表的一个实例，是对一批墓葬的属性的描述。表中的每一行记录一个墓葬的诸属性的取值，每一列记录一个属性在所有墓葬中的取值。

表 2-1　某墓地各墓葬情况的描述和随葬品统计

墓葬号	墓区	墓主人性别	墓穴大小	墓道长度(米)	器物 A 数量	器物 B 数量	器物 C 数量	器物 D 数量	随葬品总数
1	A	1	3	2	3	1	4	3	11
2	A	0	2	1	4	2	3	3	12
3	B	0	3	2	6	0	2	4	12
4	B	X	1	0	3	4	1	2	10
5	A	1	3	5	8	7	8	5	28
6	B	1	1	0	4	1	3	3	11
7	B	0	2	0	3	2	2	2	9
……	……	……	……		……	……	……	……	……
……	……	……	……		……	……	……	……	……
N	B	0	1	0	2	4	4	3	13

共统计了 N 个墓葬，每个墓葬选取了 8 个属性或变量(墓葬的序号不列为属性)，它们分别是：位于第 2 列的变量 1 反映某号墓葬所在的墓区，在 A 区，还是 B 区，属名称变量。变量 2，墓主人的性别是男还是女，用 1 表示男，用 0 表示女，属名称变量中的二元变量。其中第 4 号墓中遗骨缺失，无法判别墓主人性别，用一种特殊的符号“X”表示，称为缺失值。在原始数据中某个(或某几个)实体的某个(或某几个)属性无法观测，因此数据缺失的情况是经常发生的。我们不必因某实体的一个或两个属性无法观测而把该实体从研究对象中舍弃，在原始数据统计表中可以先用一个特殊的符号来表示，以后在处理分析这批数据时，有一系列的方法来处理缺失值。变量 3，墓穴的大小，这里用 3 表示大墓，用 2 表示中墓，用 1 表示小墓，属于有序变量。变量 4 为墓道长度，属数值变量。变量 5 至 8 统计了 4 种器物在每个墓葬中的数目，它们都是数值变量。表中的第 10 列统计每个墓中四类随葬器物的总数。其实它是第 6 至第 9 列各数值之和，是一个派生的数值，在原始数据统计表中可列可不列。总之原始数据表的每一行反映某个实体各属性的取值，是对实体的描述；而每一列给出某个属性对所有实体取值情况，反映变量取值的分布情况。

这张原始数据统计表可以方便地帮助观察一系列的问题，如随葬品的多寡是否与墓主人的性别有关，是否与墓葬的大小或与墓区有关。随葬品的多寡是随机的，还是存在个别墓葬其随葬品特别多的特殊情况。也可以研究某两个或某三个随葬器物之间是否有关联，即寻求是否存在相对稳定的器物组合，还可以按照随葬器物的情况对墓葬进行分期等。

前文提到对中原地区从二里头期到人民公园期的十多件陶豆进行分析工作，对每个陶豆的形态用了陶豆的“通高”、“口径与通高的比值”等 6 个变量来描述。其结果也是列在一张原始数据统计表中的(见第十六章，表 16-10)，并以这张表的数据作为出发点对这

批陶豆进行分期研究的。

其实原始数据统计表在传统考古学中也是常用的。在考古报告中很多资料就是用表格的形式发表的。可能有些表格中还夹杂着一些文字描述，这也是无妨的。但如果要对文字描述的内容与其他属性合在一起研究，则还需要对这些文字描述内容作定量处理。表格也是一种表达思想的语言形式，但它比用通常的文字表述有简单明了的优点。表2-1也可以用通常的文字表述，横向第一行应读作：第一号墓在A区，墓主人是男性，属大墓，墓道长度为2米，4种器物各出了多少件。第二行表述第二号墓的情况……。该表也可以竖向读：第一列表述A区和B区各有哪些墓葬，第二列表述……。通常的文字表述显然冗长噜苏，不如表格表述简明。表格表述的另一优点是把数据资料按一定的规则列在一起，有时对表格的初步观察就能发现数据中隐含的一些规律。原始数据统计表可以用计算机中常用的电子表格软件来建立，例如微软公司Office软件中的Excel电子表格软件。对于考古资料的定量研究，用电子表格软件来建立原始数据统计表是至关重要的，因为各种分析处理数据的软件都可以直接调用电子表格资料，而且电子表格本身也可以对表中的数据进行简单的统计运算，如计算极值、平均值和标准差等，可以对表中的研究对象（考古实体）进行各种各样的分组、排序等，此外电子表格还能够把表中的数据用图形的形式表达出来，从而观察数据中所隐藏的规律。

最后有两点实用的提示：(1)建立了原始数据的电子表格文件后必须仔细检查数据记录有没有错误，因为这是基础资料数据，后面要进行的数据分析处理都是建立在这张表格的基础之上的。(2)原始数据的电子表格文件应妥善保存，因为在计算机数据处理过程中，电子表格是很容易被改写的，而基础数据是不应随便被改动的。

第三章　考古资料的描述性统计

考古学研究中经常需要对一批同层次考古实体的某个属性的观察测量数据进行统计分析。例如，统计分析某墓地人骨的年龄，某地区某时代段聚落的面积，一批青铜器的含锡量或者一批砍砸器的重量等。这些人骨年龄、遗址面积、含锡量或石器重量等实际观测结果，构成一个单参数的数据组，写成 $\{X_1, X_2, \cdots\cdots X_i, \cdots\cdots X_n\}$，下标“$i$”表示数据的编号，$n$ 表示该数据组中数据的数目。这样的数据组称为由 n 个实体组成的样本，n 就是样本的容量。样本是同类实体的集合，这个概念在本书的下面章节将经常使用。单参数的数据组，或单参数的样本实际上就是第二章中原始数据统计表（表 2-1）中每一列数据的集合。

对这类数据组数据进行简单的分析统计，往往能揭示出某些简单的但可能是重要的关于考古现象中存在的规律。例如在研究聚落面积的例子中，考古学家会希望了解这些聚落按其面积的大小是怎样分布的，聚落的平均面积有多大，聚落面积间的相互差异有多大等。对聚落面积样本数据的整理，就能反映出本地区该时代段聚落面积的分布规律，就能与其他地区或其他时代段的聚落面积数据进行比较，从而进一步探讨人口与社会结构随空间和时间的变化等。又例如在研究分析某一类青铜器的含锡量时，考古学家必然会关心含锡量的代表值和分布范围等因素，这样才能进一步与其他类别青铜器的锡含量作比较，检验各类青铜器物的锡含量是否符合“六齐”之说等。相应地就要研究由聚落面积和锡含量观测值所组成的数据组的（1）数据的分布情况，（2）数据的代表值或中心值，（3）数据相对于代表值或中心值的离散程度。

对一组数据的分布、中心值和离散程度的观察分析，被称之为对数据资料的描述性统计。描述性统计往往是考古资料定量分析的第一步，很多进一步的定量研究都是建筑在描述性统计的基础之上的。

3.1　考古样本中实体的次数分布表和分布图

样本中实体的分布是指一组实体相对于其某个属性观测值的分布或分配。例如表 3-1 记录了青海乐都柳湾墓地成年女性人骨数目按年龄段的分布。

表 3-1　柳湾墓地成年女性人骨按年龄段的分布表

年龄段	青年	壮年	中年	老年	总数
人骨数	18	24	39	11	92
百分数%	19.6	26.1	42.4	12	100
累积百分数%	19.6	45.7	88.1	100	

表中的第二行反映实体按年龄段的次数分布，也称频数分布或频次分布。第三行是

百分值,称为频率分布。因为年龄段是有序变量,可将第三行的百分值累计相加。第四行是累积百分数。

实体的分布除用表格表示外,也可用图来表示,而且分布图往往比分布表能更直观地显示出实体分布的规律。常用的图形有直方图、圆瓣图、折线图和后面3.3.3节中将介绍的箱点图等。图3-1a、3-1b、3-1c分别用前三种图形表示柳湾墓地成年女性人骨按年龄段分布的情况。图3-1a是频数分布的直方图,其横坐标从左到右表示从青年到老年,纵坐标显示各年龄段的人骨数。如果纵坐标用相应年龄段人骨数所占的百分数来刻度,称为频率分布直方图。因为年龄段是有序变量,直方图上每段的宽度可以是任意的,有的书上把有序变量的直方图称之为长条图。图3-1b是圆瓣图,它反映频率分布,每个圆瓣、或扇形面积的大小正比于相应组段的百分数,各扇形面积的总和组成一个圆。图3-1c是百分累加折线图,反映某年龄段以前死亡的女性人骨的累计百分数(含相应年龄段)。

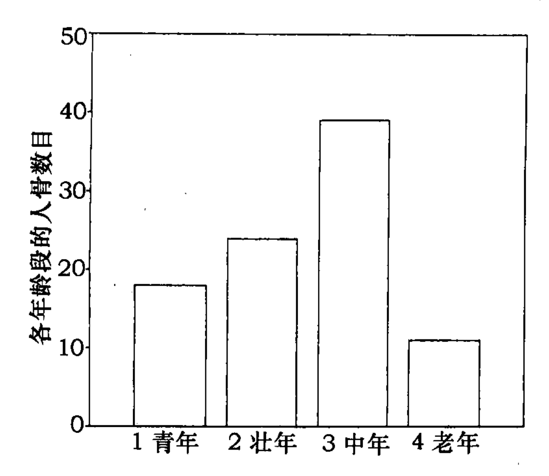


图3-1a 柳湾墓地成年女性人骨数按年龄段的分布图(a)频数直方图

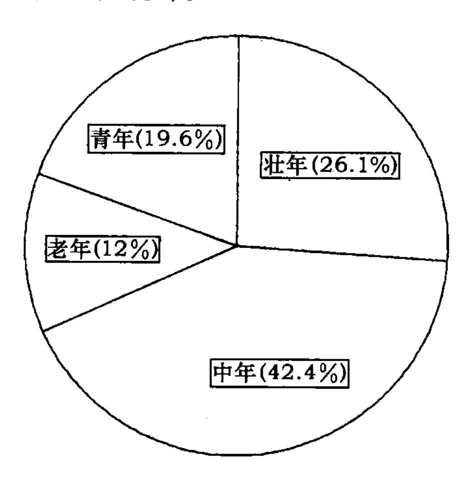


图3-1b 柳湾墓地成年女性人骨数按年龄段的分布图(b)频率圆瓣图

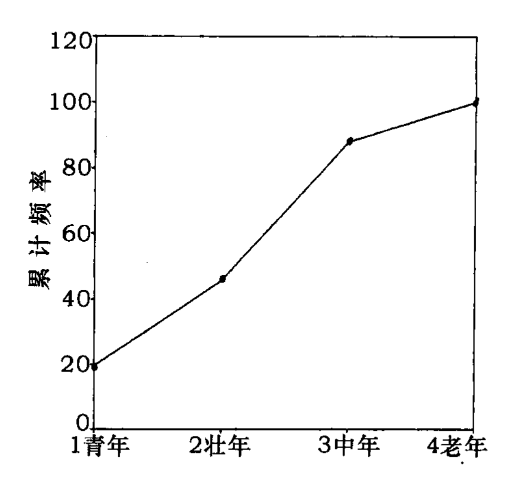


图3-1c 柳湾墓地成年女性人骨数按年龄段的分布图(c)百分累计频率图

柳湾的例子是考古实体按有序变量的分布。下面考察实体按数值属性取值的分布。英国 Dorset 地区晚新石器时代巨石文化纪念性建筑物的 35 个石柱柱洞直径的测量值统计如下(从细到粗排列,单位为 cm):25, 27, 28, 30, 34, 35, 38, 38, 38, 39, 40, 40, 40, 42, 43, 43, 43, 44, 45, 47, 47, 47, 48, 48, 48, 48, 48, 49, 50, 50, 53, 57, 57, 58, 66 (Wainwright, 1979)。这里直径是数值变量,可连续取值。为了建立柱洞数目按直径的分布表,先要对直径值分段,每段的宽度当然应该是相等的。分段的范围不同,分布表也有些差别。表 3-2a 和 3-2b 分别是取 10cm 和 5cm 为段的分布表,相应分成 5 段和 9 段。

表 3-2a　Dorset 地区巨石文化 35 个石柱柱洞按直径测量值的分布表(以 10 cm 为段)

直径范围 cm	20—29	30—39	40—49	50—59	60—69	总数
数目	3	7	18	6	1	35

表 3-2b　Dorset 地区巨石文化 35 个石柱柱洞按直径测量值的分布表(以 5cm 为段)

直径范围 cm	25—29	30—34	35—39	40—44	45—49	50—54	55—59	60—64	65—69	总数
数目	3	2	5	8	10	3	3	0	1	35

从表 3-2 可以看出,多数柱洞的直径在 40—49cm 之间,离这个中心范围愈远,柱洞数愈少,而且粗细两端的分布基本对称。为了建立分布表,并能从中较为容易地观察到柱洞数按直径分布的规律,有一个技术性的问题,即应该取多少厘米作为分段单位,或者说应把柱洞的全部直径范围分成几个等份合适。分段或分组数太少,不易分辨出分布的细致规律;分组太多,每组的实体数会很少,甚至有的段中没有实体,同样不能显示出分布的规律性。一般来说如果样本的实体数多,相应可以多分几组。作为一种实用的方法,可以以样本中实体数目的平方根作为分组数目。对于 35 个柱洞,35 的平方根值约为 6,那么大致可分为 6 组左右。因此看来前面分 5 组与 9 组都是合适的。当然分成的组段应该是互相排斥,而且组成完备的集合,即每个实体都能分到某一组而且只能归属于该组。

分布情况也可用图来表示。图3-2是Dorset石柱的柱洞按直径的频率分布直方图。

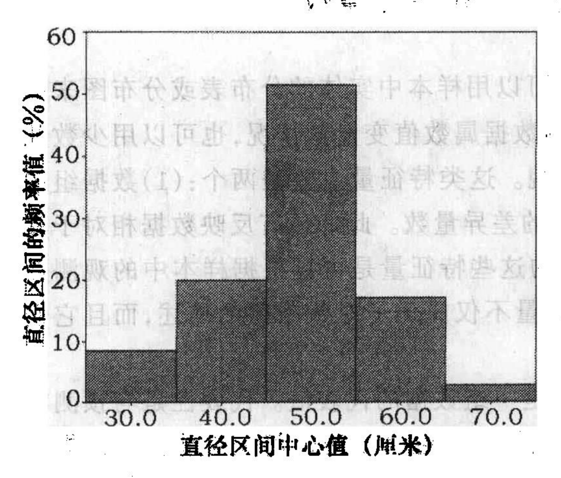


图 3-2a　Dorset 地区巨石文化石柱柱洞直径测量值的频率分布直方图(a)以 10cm 为间隔

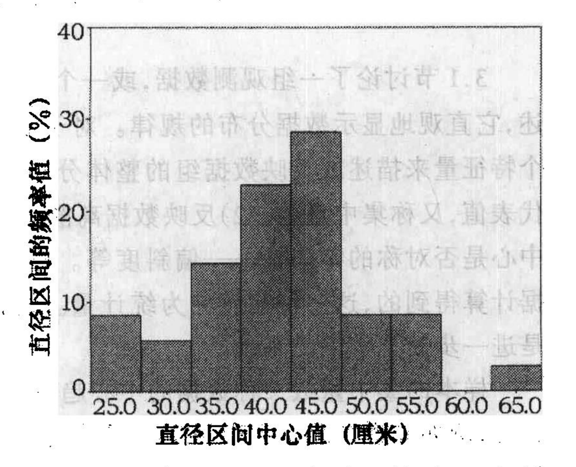


图 3-2b　Dorset 地区巨石文化石柱柱洞直径测量值的频率分布直方图(b)以 5cm 为间隔

图中横坐标表示直径值,横轴上每段的宽度都是相等的,并正比于直径段的取值范围,图3-2a和图3-2b分别以10cm和5cm为分段宽度。纵坐标表示每直径段中柱洞数所占的百分数。对于频率分布,还可以将每段的纵坐标值被每段的宽度去除,得到的商值是单位宽度的百分数或单位宽度的频率,这个"商值"称为频率密度。第四章将专门讨论频率密度的分布。

另有一种与直方图相似的显示实体分布的图形显示方法,称茎叶图(stem-and-leaf plot)。茎叶图3-3与直方图3-2b显示的内容是相似的。茎叶图用直径值的十位数(本图以每5cm为单位)组成茎,并作为茎上节点之间的间隔;而叶是由直径数值的个位数组成,在茎的相应节点上水平方向向右(或向左)生长。茎叶图比直方图保留更多的信息,因为个位数的数值也显示在图上。

茎	叶
2	5 7 8
3	0 4
3	5 8 8 8 9
4	0 0 0 2 3 3 3 4
4	5 7 7 7 8 8 8 8 8 9
5	0 0 3
5	7 7 8
6	
6	6

图3-3 Dorset地区巨石文化石柱柱洞按直径分布的茎叶图(5cm间隔)

各种统计软件,例如常用的统计软件SPSS("用于社会科学的统计软件包")等,一般都能把这些分布图画出来。

3.2 样本中数据的代表值,集中量数

3.1节讨论了一组观测数据,或一个样本可以用样本中实体的分布表或分布图来描述,它直观地显示数据分布的规律。对于观测数据属数值变量的情况,也可以用少数几个特征量来描述或反映数据组的整体分布面貌。这类特征量主要是两个:(1)数据组的代表值,又称集中量数;(2)反映数据离散程度的差异量数。此外还有反映数据相对于其中心是否对称的特征量——偏斜度等。样本的这些特征量是可以根据样本中的观测数据计算得到的,这些特征量称为统计量。统计量不仅简化了对数据组的描述,而且它们是进一步处理数据的基础。

样本的集中量数又称数据的中心趋势,它是一组数据的代表值。代表性是与预测有关的,例如知道了北京大学男生的平均身高,我们用此值去预测任何一位北大未知男生的身高,应该比用其他数值去预测最接近真实,误差最小。集中量数可以用多种方法来定义,最常用的统计量是样本的算术平均值,或简称平均值、均值,此外还有中位数(或称中数)和众值。

3.2.1　样本平均值的定义和计算

假设有样本或数据组为 $\{X_1, X_2, \cdots\cdots, X_n\}$，其平均值 $\bar{X}$ 定义如下

$$\bar{X} = (X_1 + X_2 + \cdots\cdots + X_n)/n = \sum_i X_i/n \tag{3-1}$$

用上式计算 3.1 节 Dorset 地区巨石文化 35 个柱洞直径的平均值，即将 35 个直径值求和，再除以 35，可得 $\bar{X} = 43.8$ 厘米。柱洞直径的原始测量数据用两位有效数字表示，而平均值用了三位有效数字。

当数据组可以分组时，各分组都可计算各自的平均值 $\bar{X}_i$，而且这些分组平均值以各分组样本的容量（频次）n_i 或频率 f_i 为权的计权平均正好是全组的总平均值，如下面公式所示。

$$\bar{X} = \sum_i n_i \bar{X}_i/n = \sum_i f_i \bar{X}_i \tag{3-2}$$

表 3-3 显示以 Dorset 地区巨石文化 35 个石柱柱洞直径为例分组求平均值的情况。表的最后一行第五列的单元格给出全部石柱直径的总平均值为 43.8 厘米，和公式(3-2)的计算结果是一致的。有时为了方便，求数据总平均值时不去计算各分组的平均值，而用各组取值范围的中心值 M_i 替代各分组的平均值。表 3-3 的最后一行最后一列的单元格给出了这种替代后计算的结果 $\sum_{i=1}^{5} f_i M_i = 43.07$ 厘米，与平均值的准确值 43.8 厘米稍有偏离。当数据容量 n 很大，分组可更细，而每组区间很窄时，偏离值会变得很小。

表 3-3　Dorset 地区巨石文化 35 个石柱柱洞直径按分组求平均值

直径范围 cm	频次 n_i	频率 %f_i	组平均值 $\bar{X}_i$	$f_i \bar{X}_i$	组中心值 M_i	$f_i M_i$
20—29	3	8.57	26.67	2.29	24.50	2.10
30—39	7	20.00	36.00	7.20	34.50	6.90
40—49	18	51.43	45.00	23.14	44.50	22.89
50—59	6	17.14	54.17	9.29	54.50	9.34
60—69	1	2.86	66.00	1.89	64.50	1.84
列的总和	35	100.00		43.80		43.07

3.2.2　中位数和其他的集中量数

另一个常用于描述样本的集中量数是中位数，或中数，也称中值。其定义是将数据组的数据按大小次序排列好后，该序列中央的那个数。中数前面的数据数目和其后面的数据数目正好相等。3.1 节中 35 个柱洞直径值的中数是 44 厘米，它是 35 个柱洞按照它们的直径值从小到大排列中的第 18 个柱洞的直径值。44 厘米与平均值 43.8 厘米非常接近。这里柱洞数为 35，是一个奇数，如果样本中数据的数目是偶数，可以先找到中央的两个数据，再取这两个数据的平均值作为该数据组的中数。

还有一个集中量数是众数，它是由数据组中出现次数最多的那个数来决定的。35 个柱洞直径值中出现最多的数是 48，它出现了 5 次，因此这组数据的众数是 48，它与平均

值、中数都有一定差距。只有当样本的容量非常大,即数据组数据数目很多,而且实体的分布接近后面将介绍的正态分布时,众数才有意义,其数值会与平均值和中数非常接近。当然原则上也可以用样本的几何平均值、调和平均值等统计量作为其集中量数的指标,但应用甚少,这里不作讨论。

3.2.3 平均值和中位数的比较

对于数值变量,样本最常用的集中量数是平均值,因为平均值有严格的数学定义,它是很多统计分析方法的基础,而且平均值概念在日常生活中也被广泛使用,容易被理解和接受。平均值作为样本的代表值的缺点是,当样本容量不太大时平均值的稳定性不如中数,它受极端数据的影响较大。中数的优点是它表达较低一半数据和较高一半数据的界限,受极端数据的影响很小。例如有一组数据 {2, 2.4, 2.5, 2.7, 3.0, 3.1},它的平均值和中数分别为 2.62 和 2.6 。如果数组中加进一个极端值 10,平均值变成 3.67,而中数变化却不大,取值为 2.7。中数还可以应用于有序变量。但是在样本可以分组情况下,各分组的中数与全组的中数间不一定有什么关系。为了降低极端数据对平均值的影响,可以计算 5%剪裁平均值(5% trimmed mean),它是排除样本中偏离平均值最远的 5%数据后重新计算的平均值。剪裁的标准,即被排除的极端数据的百分数是可以变动的。

一个需要注意的问题是:如果一组数据的分布呈双峰分布,那么该数据组的集中量数,无论是平均值或是中数,都是没有意义的。例如一个样本中包括长剑和短佩剑两种不同类型的剑,计算样本中剑的平均长度是没有意义的。又例如托儿所里有身高 1.6 米左右的老师阿姨,也有身高不足 1 米的儿童,求某托儿所全体人员的平均身高或中数是毫无意义的。应分别考虑老师阿姨的平均身高和儿童的平均身高。因此在计算一组数据的平均值时,应先检查一下数据的分布,观察是否为单峰分布,也就是说要确认我们所研究的样本中的实体应该属于同一类型的,否则求样本的集中量数是没有意义的。

3.3 样本中数据的离散程度、差异量数

一组数据的分布特征仅用平均值或中数等集中量数来表征是不完善的,还必须注意组内各数据之间的离散程度。例如有两个数组,分别是:{1, 1.5, 1.5, 2, 2.5, 2.5, 3} 和 {1.8, 1.8, 1.9, 2, 2.1, 2.2, 2.2},它们的平均值是相等的,都等于 2,它们的中数也都等于 2。但可以看出,这两个数据组的数据分布的离散程度是不一样的,相比之下第二个数组中的数据较为集中,都离中心“2”不远。

3.2 节讨论的集中量数反映的是一组数据的代表值。如果一组数据是记录某类陶器的线性尺度,其集中量数应是设计的尺寸,那么数据的离散程度反映陶器加工的工艺水平,陶器的实际制作在多大程度上符合原设计的指标。专业加工的陶器其实际尺寸的离散性小,而家庭作坊加工的产品就不那么规范,产品的尺寸离设计值会有较大涨落。因此需要定义表征样本的数据间离散程度的量,称为差异量数,它也是一个重要的统计量。常用的描述数据间离散程度的统计量是标准差和四分位差,但标准差必须与平均值一起使用,而四分位差与中数组成一对指标来描述数据组的集中量数和差异量数。

3.3.1　样本方差和标准差的定义和计算

对一组数据（X_1，X_2，……X_i，……X_n）计算出平均值 $\overline{X}$ 后，可以算出数据组中每个成员与平均值的差：

$$x_i = X_i - \overline{X} \tag{3-3}$$

x_i 称为离差。离差的值可以为正，也可以为负。根据对平均值的定义，样本的离差和 Σx_i 总是等于零的。在统计学中是用方差 S^2 和标准差 S 来表示数据间的离散程度。S^2 与 S 是通过这些离差 x_i 计算得出的统计量。计算公式如下：

$$S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n = \sum_{i=1}^{n}x_i^2/n \tag{3-4}$$

$$S = \sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2/n} \tag{3-5}$$

一组数据的方差是其每个成员与平均值之差的平方和(即离差平方和)再被数据的数目除，即是平均离差平方和。而标准差是方差的平方根值。方差 S^2 也可以用下面的公式(3-6)来计算，在计算机没有普及的年代，这是一个比较简易地计算方差值的公式。

$$S^2 = \overline{X^2} - (\overline{X})^2 \tag{3-6}$$

由公式(3-6)可见，数组的方差等于该数组各元素平方的平均值减去数组平均值的平方。公式(3-6)证明如下：

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + (\overline{X})^2) \\ &= \frac{1}{n}\Big(\sum_{i=1}^{n}\big(X_i^2 - 2\overline{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}(\overline{X})^2\Big) \\ &= \overline{X^2} - 2\overline{X}\,\overline{X} + (\overline{X})^2 = \overline{X^2} - (\overline{X})^2 \end{aligned}$$

根据上面两个公式可以计算出 Dorset 地区巨石文化柱洞直径的方差为

$$S^2 = \sum_{i=1}^{35}(X_i - 43.8)^2/35 = \sum_{i=1}^{35}x_i^2/35 = 79.2$$

单位是 cm^2；而标准差是 $\sqrt{79.2} = 8.90$，单位用 cm 表示。本节{1，1.5，1.5，2，2.5，2.5，3} 和 {1.8，1.8，1.9，2，2.1，2.2，2.2}两组数据的标准差分别为 0.65 和 0.16。第一组数据的数据取值分散，因此标准差比第二组数据大，显示了标准差反映了数据的离散程度。在下面章节中将介绍，很多情况下数据的分布接近所谓的“正态分布”，这时大约有 68.3 % 的数据会处于以平均值为中心，二倍标准差为宽度的区间 [$\overline{X} \pm S$] 中。

3.3.2　总体标准差和样本标准差

在统计学中总体和样本是一对十分重要的概念，这在以后的章节中会详细讨论。这里先指出，公式(3-5)和(3-6)是计算总体标准差的公式，计算样本标准差的公式应该是：

$$s = \sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)} \tag{3-7}$$

样本的标准差我们用小写的 s 表示，计算公式的分母上是用 $(n-1)$ 取代了公式(3-4)和(3-5)的分母上的 n。利用公式(3-7)计算 Dorset 柱洞直径样本的标准差 $s=9.03\text{cm}$，大于 $S=8.90\text{cm}$。s 比 S 略大，但当样本容量 n 很大时，这两个量的数值是十分接近的，其数值上的差别可以忽略不计了。

还有一个反映一组观测值离散程度的量称为相对标准差或变异系数 η。变异系数定义为：

$$\eta = s/\bar{X} \text{ 或 } \eta = S/\bar{X} \tag{3-8}$$

变异系数用百分数表示。Dorset 柱洞直径数据组的变异系数为 $9.03 \div 43.8 = 20.6\%$。

3.3.3 四分位数和四分位差

3.3.2 节定义的标准差是与平均值搭配使用的，如果用中数当作数据组的代表值，则反映离散程度的差异量数是四分位差。为此先定义四分位数，四分位数是按大小排列的数组中处于四分之一和四分之三位置上的两个数据，分别称为上、下四分位数，用 Q_{25} 和 Q_{75} 表示。$(Q_{75}-Q_{25})$ 是两个四分位数的差，又称四分位差，数组中有 50% 的数据落在这个区间中。

我们仍以 Dorset 柱洞直径样本为例来说明四分位数和四分位差。重新抄录 Dorset 的数据如下：

25，27，28，30，34，35，38，38，**38**，39，40，40，40，42，43，43，43，**44**，
45，47，47，47，48，48，48，48，**48**，49，50，50，53，57，57，58，66

可见 Dorset 柱洞直径样本的 Q_{25} 和 Q_{75} 为第 9 位数据“38”和第 27 位数据“48”，四分位差为 $48-38=10$。另外 Q_{50} 就是中数，而 Q_0 和 Q_{100} 是数组的最小和最大两个极值。一般情况下，如果样本的容量为 n，那么 Q_{25} 的位置为 $(n+1)/4$，中值 Q_{50} 的位置为 $(n+1)/2$，Q_{75} 的位置为 $3(n+1)/4$。

3.3.4 反映数据分布的箱点图

在 3.1 节中曾利用直方图和茎叶图来显示 Dorset 柱洞直径数据的分布，另一种显示数据分布的常用方法是箱点图(Box-and-dot plot)，也称箱图或 Box-Whisker 图。图 3-4 是 Dorset 石柱直径分布的箱点图表示。

箱点图是以两个四分位数(38 和 48)为界做一个箱体，箱体的高度就是四分位差($48-38=10$)，50% 的数据落在箱体的区间中。在代表中数数值的位置(44)处也画一水平线段，该线段接近于箱体中央，但不一定处于箱体的正中央。下一步是确定邻近区域和特殊歧离点。作为一种约定俗成的规则，离箱体的上下边缘以箱体高度的 1.5 倍为距离作为标准。箱体边缘至这两个标准值之间的区间称为临近区域，取值超过这两个标准的实体被认为是明显偏离样本中心的特殊歧离实体。在 Dorset 石柱的例子中，这两个标准分别为 $48+1.5\times10=63\text{cm}$ 和 $38-1.5\times10=2.3\text{cm}$。在 35 个柱洞直径值中其直径为 66cm 的柱洞属于特殊实体，因为 $66>63$。除去这个特殊实体外，其他实体或位于箱体中，或位于邻近区域。箱体图还规定需要分别标出邻近区域中取值最大和最小的实体的位置。标识的方法可以在这两个实体的位置处划一水平线，对于所讨论的样本这两个线段的位

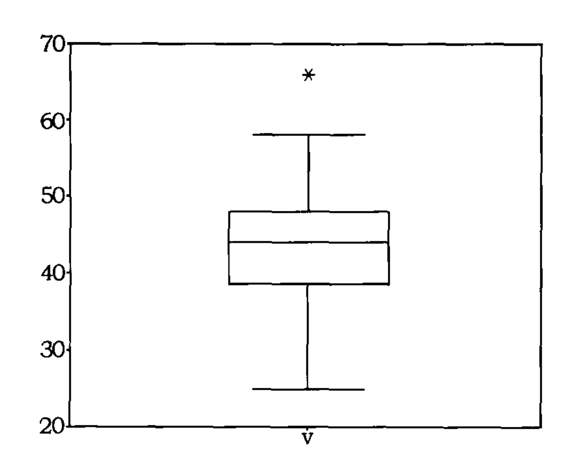


图 3-4　Dorset 石柱直径分布的箱点图表示

置应该在 58cm(数组中第二大的数)和 25cm 处(数组中最小的数)。数据组中除特殊歧离点外,其他的数据点均落在 58cm 和 25cm 这两条线段之间。与 3.1 节的直方图和茎叶图相比,箱点图明确显示了数据中心值的位置,中间 50%数据的位置,全部非特殊数据的分布范围,并给出了判断特殊偏离点的一种标准。因此箱体图在说明性数据分析的方法中很被推崇。与箱点图相似的还有一种称之为子弹形图的表示方法,子弹形图能直观地应用于样本间平均值的比较,它是基于对平均值估计区间置信度的概念。本书在第八章的 8.4 节将予以介绍。

3.3.5　标准差和四分位差的比较

方差和标准差有严格的数学定义,是概率统计学的基础概念,在本书后面要讨论的数值变量的总体参数估计,平均值的假设检验,相关和回归乃至本书下篇介绍的各种多元统计方法都涉及这两个概念。但标准差的缺点和平均值相似,受偏离大的极值的影响大。相对而言,四分位差比较稳定,不受或少受偏离大的极值的影响。根据四分位差建立的箱点图能直观地显示数据点的分布范围和特殊点。此外四分位差也可应用于有序变量的情况。

20 世纪 70 年代起 W. Tukey 提倡一种称为说明性数据分析的方法(Exploratory Data Analysis ,简称 EDA)。提倡使用茎叶图、箱点图、中位数和四分位差等方法和概念来分析数据。EDA 方法的优点是减少了复杂的数学计算,分析结果直观,容易理解。随着一些统计软件包逐步将 EDA 的各种方法的纳入,EDA 方法的使用在各领域,包括考古资料的分析,也渐趋普及。例如匹兹堡大学 R. Drennan (周南)编写的《Statistics for Archaeologists》教科书中介绍 EDA 方法的篇幅占了相当的比例。

3.4　EXCEL 软件应用于数据组的描述性统计

本章所讨论的平均值、标准差、中数和四分位数等在 Excel 软件中均可用相应的函数计算。这些函数依次是 average (数组)、stdevp (数组)或 stdev (数组)、median(数组)和

quantile（数组,0－4）。这里 stdevp 和 stdev 分别计算总体标准差和样本标准差,视情况选用。计算四分位数的 quantile（数组,0－4）时,除需输入数组外,还要对后面的开关赋值,0 到 4 分别计算 Q_0 到 Q_{100},即从数组的最小值、上四分位数、中数、下四分位数到最大值。使用 SPSS 软件也可以方便地计算这些统计量,第十三章专门对此作介绍。

第四章　考古统计学的基础知识准备
——概率基础知识和两个重要的理论分布

4.1　概率基础知识复习

考古学以古代人类活动所留下的遗迹遗物作为研究对象。但是遗迹遗物能否被保存下来，又能否被考古学家所发现和发掘有很大的随机性。概率论作为数学的一个分支，专门处理随机现象。概率的研究最早起源于研究赌博，因为投掷骰子和玩扑克牌充满了随机现象。据传罗马皇帝克迪斯一世（公元前 10—54）曾撰写了一本名为“赌赢秘诀”的书，很可惜失传了。文艺复兴时期有一位梅雷爵士请他的朋友，著名的法国数学家、物理学家布莱茨·珀斯卡（1623—1666）解一道骰子赌博的难题。“1 个骰子抛掷 4 次至少一次是 6 点，和 2 个骰子抛掷 24 次至少一次是双 6 点；哪种机会更多”？为了解决这个问题，帕斯卡与费马进行了讨论，后来俩人共同奠定了概率论的基础。有了概率论的知识，梅雷爵士提出的问题是不难回答的，本节后面将给出答案。为了便于读者的理解，本节将用抛掷骰子和扑克牌的例子来介绍或者复习有关概率论的一些基本概念。

4.1.1　概率的定义

什么是概率，概率与第三章中介绍的频率的概念是紧密相关的。第三章的表 3-1 统计乐都柳湾墓地 92 具成年女性人骨中有 18 名是青年女性的人骨，从而计算出青年女性人骨出现的频率为 $18 \div 92 = 19.2\%$。柳湾墓地发掘出的成年女性人骨的数量是有限的，所定的频率值 19.2%并不能精确地代表墓地所属氏族青年女性的死亡频率，而只是一个近似值。在柳湾墓地随意找一个女性人骨，并不能完全准确地预测她属哪个年龄段，而只能给出她属于哪个年龄段的大致可能性。我们再举一个投掷骰子的例子。投掷 6 次，出现“4 点”的次数不一定是 1 次，可能一次也不出现，也可能出现 2 次，甚至 3 次。就是说出现“4 点”的频率不一定是 1/6。而且即使最初的 6 次投掷“4”出现 1 次，计算得到频率值是 1/6，但再继续投掷下去，出现“4”的频率还是会偏离 1/6 的。但是如果增加投掷次数，譬如投掷 60 次，甚至 600 次，出现“4 点”的频率会愈来愈接近 1/6，偏离愈来愈小。我们可以把上面的例子归纳如下：事件 A 在每次试验中是否出现有一定的偶然性的。一定的条件下进行了 n 次试验，其中事件 A 出现了 m 次，那么事件 A 出现的频率为 $f\{A\} = m/n$。在相同的条件下再进行了 n 次试验，事件 A 出现的次数就不一定还是 m 次，但也不会偏离 m 太远。试验的次数 n 愈多，频率 m/n 的数值愈稳定。这种稳定性，或规律性称作客观的统计规律性。当试验次数非常非常多时，事件 A 出现的频率 $f\{A\}$ 趋向一个确定的数值，称为事件 A 在每次试验中出现的概率 $P\{A\}$。写作

$$P\{A\} = \lim_{n \to \infty} f\{A\} \tag{4-1}$$

不拘于数学上的严格性，上面我们用频率法来理解和计算概率，把某感兴趣事件 A 出现的概率理解为在一次试验中该事件出现的可能性。

我们也可以通过一类简单又常见的，称为古典概型的随机现象来理解概率。古典概型随机现象的特点是：(1)试验结果的个数是有限的（n 个）；(2)各种试验结果出现的可能性是相等的。例如一付扑克牌任抽取一张，只可能有52种结果，而每一张牌被抽取的可能性是相等的。这52种可能的抽取结果称为52个互不相容的基本事件。由于只可能有这52种结果，它们又组成了完备的基本事件组，写成 $U=\{U_1,U_2,\cdots\cdots,U_n\}$。每次试验中每个基本事件出现的概率就应该是 $P\{U_i\}=1/n$ 而且有

$$\sum_{i=1}^{n} P\{U_i\} = 1 \tag{4-2}$$

公式(4-2)表示 $U=\{U_1,U_2,\cdots\cdots,U_n\}$ 组成了完备的互斥的基本事件组。在处理古典概型的随机现象时，每个事件 A，都可以看成由若干个(例如 m 个)基本事件组成的。在抽扑克牌的试验中，抽取一张牌是红桃的事件是由抽取红桃 A 到红桃 K 这13个基本事件所组成。因为基本事件是互不相容的，事件 A 的概率定义为

$$P\{A\} = m/n \tag{4-3}$$

抽取一张牌是红桃的概率就等于 13/52 = 1/4。

4.1.2 概率运算的基本法则和应用实例

为了计算一些更复杂事件的概率，下面讨论概率运算的几个法则。

1. 加法法则。现有事件 A 和 B，新事件 C 由"A 事件发生或 B 事件发生(当然也包括 A，B 同时发生)组成。我们称 C 是 AB 事件的和，记作 $C=A\cup B$。举例来说抽一张扑克牌或者是红桃或者是"K"都可以接受。事件和 C 的概率是

$$P\{A\cup B\} = P\{A\} + P\{B\} - P\{A\cap B\} \tag{4-4}$$

式中 $A\cap B$ 表示 A 与 B 同时发生的事件，称为 AB 两事件的积，而 $P\{A\cap B\}$ 是 A 与 B 同时发生的概率。因为在 $P\{A\}$ 和 $P\{B\}$ 中都包含了 A 与 B 同时发生的概率，在计算 $P\{A\cup B\}$ 时不应重复计算，因此在公式(4-4)中要扣除一项 $P\{A\cap B\}$。利用加法法则可以计算出，抽一张牌或者是红桃或者是"K"的概率应该等于：

$$P\{C\} = (13/52 + 4/52 - 13/52 \times 4/52)$$

加法法则的一种特殊情况是，如果 A 与 B 是互不相容的事件，即它们不可能同时发生，这时 $P\{A\cap B\}=0$，公式(4-4)改写成

$$P\{A\cup B\} = P\{A\} + P\{B\} \tag{4-5}$$

还是以抽取扑克牌为例子，抽红桃和抽方片是互不相容的事件，抽一张牌是红色的概率就是抽红桃和抽方片两个互不相容的事件的概率之和，为 13/52 + 13/52 = 0.5。

2. 乘法法则。现有事件 A 和 B，新事件 C 是"A 事件和 B 事件同时发生"，称为 AB 的事件积，写作 $C=A\cap B$。例如要求从一副扑克牌中先抽一张是红桃(事件 A)，放在一边，再抽第二张还是红桃(事件 B)。事件积的概率为

$$P\{A\cap B\} = P\{A\}P\{B\mid A\} \tag{4-6}$$

式中 $P\{B\mid A\}$ 称为事件 A 已发生的情况下发生事件 B 的条件概率。这样连抽两张红桃的

概率为(13/52 × 12/51)。因为一副牌中已抽走一张红桃，只剩下51张牌，其中仅有12张红桃，因此抽第二张还是红桃的条件概率为12/51。

有一种特殊情况是事件 A 和 B 是相互独立的，即事件 B 的发生与 A 事件是否已先发生无关，这时条件概率 $P\{B \mid A\} = P\{B\}$，事件积的概率的计算公式成为

$$P\{A \cap B\} = P\{A\}P\{B\} \tag{4-7}$$

还是以连抽两张扑克牌为例，但抽取的条件改为，抽取第一张后放回去，再抽第二张，这时抽取第二张的概率与第一次的抽取结果无关，这种情况下连抽两张都是红桃的概率为(13/52 × 13/52)。

3. 减法法则。事件 A 可以发生，也可以不发生，我们把不发生事件 A 称之事件 A 的逆事件，或非 A 事件，记作 $\bar{A}$。A 与 $\bar{A}$ 互不相容，且组成完备的事件组，因此有：

$$P\{A\} + P\{\bar{A}\} = 1 \qquad 或 \qquad P\{\bar{A}\} = 1 - P\{A\} \tag{4-8}$$

4. 全概率公式。有时直接计算某个复杂事件 B 发生的概率不太方便，可以利用全概率公式来求解。假设 A_1，A_2，……，A_n，组成完备的互不相容事件组，即有

$$P\{A_i \cap A_j\} = 0 \qquad \sum_i P\{A_i\} = 1$$

因此如果发生了 B 事件，它必定与 A_i 中的某个事件同时发生，而且只是与该事件同时发生。即 $B \cap A_1$，$B \cap A_2$，……，$B \cap A_n$ 也同样组成完备的互不相容事件组，因此利用不相容事件的概率加法法则，可以写出

$$P\{B\} = \sum_i P\{B \cap A_i\}$$

再利用计算事件积概率的公式(4-6)，就可推导得到全概率公式

$$P\{B\} = \sum_i P\{A_i\}P\{B \mid A_i\} \tag{4-9}$$

例题　一张张地抽扑克牌，抽出后不放回，求抽第三张是红桃的概率。现将求抽第三张是红桃的事件称为事件 B。因为抽出的牌不放回，抽第三张牌的概率依赖于前面两次抽取的结果。前两次抽取可能发生4种情况：

A_1　第一张红桃，第二张也是红桃

A_2　第一张红桃，第二张不是红桃

A_3　第一张不是红桃，第二张是红桃

A_4　第一张不是红桃，第二张也不是红桃

这四个事件组成了完备的不相容事件组。用全概率公式可计算抽第三张是红桃的概率。

出现 A_i 的概率	A_i 出现后出现 B 的条件概率
$P\{A_1\} = 13/52 \times 12/51$	$P\{B \mid A_1\} = 11/50$
$P\{A_2\} = 13/52 \times 39/51$	$P\{B \mid A_2\} = 12/50$
$P\{A_3\} = 39/52 \times 13/51$	$P\{B \mid A_3\} = 12/50$
$P\{A_4\} = 39/52 \times 38/51$	$P\{B \mid A_4\} = 13/50$

这样在前两张扑克牌抽出后不放回的条件下，抽第三张是红桃的概率 $P\{B\}$ 为

$$P\{B\} = \sum_i P\{A_i\}P\{B \mid A_i\}$$

$$= \frac{1}{52 \times 51 \times 50}(11 \times 12 \times 13 + 2 \times 12 \times 13 \times 39 + 39 \times 38 \times 13) = 0.25$$

这个结果与前两张扑克牌抽出后放回,抽第三张是红桃的概率是一致的。

5. 逆概率公式。全概率公式是已知诸事件 A_i 组成完备的事件组,而根据 B 事件发生的诸原因 A_i,来计算 B 事件发生的概率,是从原因来推算结果。反过来,如果 B 事件已经发生,希望来探求组成完备事件组的诸原因 A_i 各导致 B 事件发生的概率多大,是由果探因。这类概率的计算公式称为逆概率计算公式,又称为贝叶斯公式。

假设 $A_1, A_2, \cdots\cdots, A_n$ 组成完备的互不相容事件组。现在事件 B 已经发生,要计算它是由 A_i 导致的概率,即计算条件概率 $P\{A_i \mid B\}$。利用事件积概率的公式

$$P\{B \cap A_i\} = P\{A_i\}P\{B \mid A_i\} = P\{B\}P\{A_i \mid B\}$$

整理得到

$$P\{A_i \mid B\} = \frac{P\{A_i\}P\{B \mid A_i\}}{P\{B\}}$$

再将全概率公式(4-9)取代上式分母中的 $P\{B\}$,得逆概率公式

$$P\{A_i \mid B\} = \frac{P\{A_i\}P\{B \mid A_i\}}{\sum_i P\{A_i\}P\{B \mid A_i\}} \tag{4-10}$$

下面通过一个具体的例子来说明逆概率公式的应用。在某地区流行一种传染病,已知有千分之一的人得病。有一种检验方法,它对病人的检出率为 99%,但对健康人检验的假阳性率为 2%。现在有一位张先生检查为阳性,我们希望知道他已经传染得病的概率是多少。得病 (A) 和健康($\bar{A}$) 两种状态构成了完备的互不相容事件组,而且已知得病和健康的概率分别为 $P\{A\} = 0.001$ 和 $P\{\bar{A}\} = 0.999$。规定检验结果为阳性为事件 B,则 $P\{B \mid A\} = 0.99$ 和 $P\{B \mid \bar{A}\} = 0.02$。检验阳性且已经传染得病的概率是 $P\{A \mid B\}$。根据逆概率公式

$$P\{A \mid B\} = (0.001 \times 0.99)/(0.001 \times 0.99 + 0.999 \times 0.02) = 0.0472$$

就是说张先生已传染得病的概率小于 5%,需要进一步观察或隔离,但也不必过于紧张。顺便我们还可以计算这种检验方法的有效性。假设对 10000 人作了检验,那么平均而言有 $10000 \times (0.001 \times 0.99 + 0.999 \times 0.02) = 210$(人)检验阳性。其中 200 为假阳性。另一方面真正得病而未查出的人数平均为 $10000 \times 0.001 \times (1 - 0.99) = 0.1$(人)。当然0.1(人)是没有意义的,人数必须整正数,但说明 10000 人中未检出的病人是极少的。使用这种检验方法后,10000 人中仅需对 210 人做监控就可以了,而且真正病人未被检查出的风险极小。

了解了概率运算的基本法则,可以来回答梅雷爵士向帕斯卡请教的问题。第一个问题是投掷 4 次骰子至少出现 1 次 6 点的概率,用 E 表示这个事件。直接计算这事件的概率 P{E} 较复杂,我们计算 E 的逆事件,即投掷 4 次一次也没有出现 6 点的概率 $P\{\bar{E}\}$。单次投掷不是 6 点的概率是 5/6,而每次投掷都是独立事件,因此用乘法法则求投掷 4 次 1 次也没有出现 6 点的概率 $P\{\bar{E}\} = (5/6)^4 = 0.482$。根据减法法则 $P\{E\} = 1 - P\{\bar{E}\} =$

$1-0.482=0.518$,即投掷4次骰子至少出现1次6点的概率是0.518。梅雷爵士的第二个问题是24次同时投掷2个骰子,至少出现一次双6点的概率是多少。用 Y 表示这个事件。同样先计算其逆事件的概率,即24次投掷骰子未出现一次双6点的概率 $P\{\bar{Y}\}$。投一次不是双6的概率是 $1-(1/6)^2=35/36$。24次投掷未出现一次双6的概率 $P\{\bar{Y}\}=(35/36)^{24}=0.509$。24次投掷至少出现一次双6点的概率 $P\{Y\}=1-\{\bar{Y}\}=1-0.509=0.491$。现在可以告诉梅雷爵士,事件 E 出现的概率略大于出现事件 Y 的概率,他可以参考上面计算的概率值来下赌注或计算有关的赔偿率。

4.2　排列和组合知识复习

在计算复杂事件的概率时,以及后面要讨论的二项式分布时都需要一些关于排列和组合的知识,下面作简要介绍。

(一) 排列问题。

假设有 n 个不同的元素 $a_1,a_2,\cdots\cdots,a_n$ 组成一个集合,从中任意取出 m 个($m\leqslant n$),并加以排列,问有多少种排列方法。这里有两种抽取的方法。第一种是抽取出的元素要放回。这样抽取的元素可以是重复的。抽取 m 个并排列的方法的数目为

$$R_n^m=n\times n\cdots\cdots n=n^m \tag{4-11}$$

第二种方法是抽取出的元素不放回。那么抽取第一个元素有 n 种方法,取第二个有 $(n-1)$ 种方法,抽取第 m 个有 $(n-m+1)$ 种方法。这种情况下,不同的排列方法数目为

$$P_n^m=n(n-1)(n-2)\cdots\cdots(n-m+1) \tag{4-12a}$$

将 $n\cdot(n-1)(n-2)\cdots\cdots3\cdot2\cdot1$ 记作 $n!$,则公式(4-12a)改写为

$$P_n^m=\frac{n!}{(n-m)!} \tag{4-12b}$$

当全部元素都抽取时,即 $m=n$ 时,则有

$$P_n^m=n! \tag{4-13}$$

例题　4名围棋运动员选3名,并按第一到第三比赛台排列,问有几种选择方法。答案是 $P_4^3=4!/(4-3)!=24$。有24种选择方法。

(二) 组合问题。

如果从 n 个元素中任意取出 m 个而不加以排列,问有几种取出方法,称为组合数,记作 C_n^m,显然

$$C_n^m=\frac{P_n^m}{m!}=\frac{n!}{(n-m)!\cdot m!} \tag{4-14}$$

例题　在一批墓葬中鉴别出4种类型的器物,以A,B,C和D命名。如果3种类型的器物形成一种组合,求理论上有多少种可能的组合。这里 $n=4,m=3$。可能的组合数为 $C_4^3=\dfrac{4!}{(4-3)!3!}=\dfrac{4\times3\times2\times1}{1\times(3\times2\times1)}=4$。如果鉴别出的器物类型是5种,则3种类型器物

形成的组合理论上可能有 C_5^3 种，$C_5^3 = \frac{5!}{(5-3)!3!} = \frac{5\times4\times3\times2\times1}{(2\times1)\times(3\times2\times1)} = 10$。

在 Excel 软件中可以用函数 PERMUT (n, m) 和 COMBIN (n, m) 来计算 P_n^m 和 C_n^m。对上面的例题可分别键入“ = PERMUT(4,3)”和“ = COMBIN(4,3)”，将分别返回“24”和“4”。

4.3 均匀分布

第三章曾讨论了一组实体依据其某一属性的频次分布和频率分布，这是样本的经验分布。至于总体的分布，在一些情况下可以依据我们关于总体的知识，通过逻辑推理来建立，这是关于总体的理论分布。理论分布可以用一定的数学函数来表述。自本节起将依次讨论均匀分布、二项式分布和正态分布等三种理论分布。

首先讨论均匀分布。我们还是从抽扑克牌为例着手。每次抽一张，记录牌的点数后放回。每次试验的所得到的“点数”是一个随机变量，变量取正整数，并且其取值范围是从 1(A)到 13(K)变化。根据我们对扑克牌组成的知识，可知该随机变量取 13 个可能值中任何一个值的概率都是相等的，都等于 4/52 = 0.077，其概率分布是一个均匀分布，如图 4-1 所示。实际抽取扑克牌所得的频率分布属于经验分布，由于随机的涨落，抽扑克牌的经验分布不可能如此理想地均匀，但当抽取次数不断增加时，经验分布也愈益接近理论上的均匀分布。

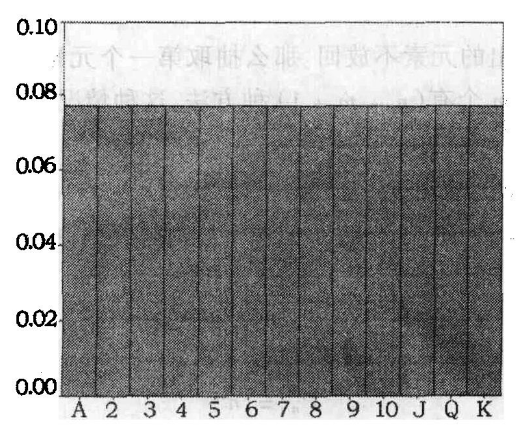


图 4-1 回放抽取单张扑克牌出现的点数的均匀概率分布图

4.4 二项式分布

4.4.1 贝努里试验和二项式分布

假设在固定条件下重复一系列独立的试验，而每次试验只可能出现 A 与非 A($\bar{A}$)两种结果，这样的试验称为贝努里试验。所谓独立的试验是指每一次试验结果的概率与先前各次试验的结果无关。例如多次的投掷骰子或硬币，从完整的扑克牌中任意抽取一张

牌，从一个墓地中每次新鉴定一个人骨的性别等都属于独立的试验。但其中只有投掷硬币和人骨性别鉴定属贝努里试验，而投掷骰子和抽取扑克牌会产生多种结果，不属于贝努里试验。贝努里试验所产生的随机变量只可能有两个取值："是或非"，称为二元随机变量。二项式分布就是二元随机变量的概率分布。

假设每次贝努里试验出现 A(成功)与非 A(失败)的概率分别为 p 和 q。每次试验的结果，成功与失败必居其一，因此有

$$p + q = 1 \tag{4-15}$$

进行 2 次试验可以出现 3 种结果：(1)2 次都成功，其概率为 p^2；(2) 一次成功一次失败，因为从结果看不必区分成功与失败的先后，这样一次成功一次失败的概率是 $2pq$；(3) 连续两次失败，其概率为 q^2。这 3 种结果之中必然出现一种，因此有

$$p^2 + 2pq + q^2 = (p + q)^2 = 1 \tag{4-16}$$

进行 3 次试验可以出现 4 种结果：(1)3 次成功，(2)2 次成功 1 次失败，(3)1 次成功 2 次失败和(4)3 次失败。下面公式(4-17)中的 4 项分别表示这 4 种结果出现的概率，以及它们的概率和为 1。

$$p^3 + 3p^2q + 3pq^2 + q^3 = (p + q)^3 = 1 \tag{4-17}$$

如果进行了 n 次试验，则可能产生 $n+1$ 种结果。且有

$$\sum_{m=0}^{n} C_n^m p^m q^{(n-m)} = (p + q)^n = 1 \tag{4-18}$$

式中 $C_n^m p^m q^{(n-m)}$ 是进行 n 次试验有 m 次成功和 $(n-m)$ 次失败的概率。C_n^m 是组合数，$C_n^m = \dfrac{n!}{(n-m)! \cdot m!}$。

从上面的分析中看到，贝努里试验结果的概率分布相似于我们所熟悉的二项式的展开，因此称为二项式分布。二项式分布有二个参数：试验总次数 n 和每次试验成功的概率 p。图 4－2 是 $n = 6, p = 0.5$ 的二项式分布图

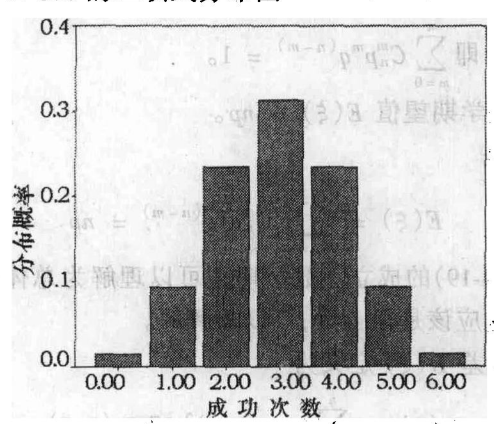


图 4-2　$n = 6, p = 0.5$ 的二项式分布图

已知对于贝努里试验，n 次试验可能产生 $n+1$ 种结果。令随机变量 ξ 的取值等于成功的次数 m，就有

$$P\{\xi = m\} = C_n^m p^m (1-p)^{(n-m)} \qquad (m = 1,2,\cdots n)$$

除了利用公式(4-18)计算贝努里试验中随机变量 ξ 取某个数值的概率外,还可以计算随机变 ξ 取值为某一范围的概率。

(1) 几次试验成功的次数不少于 r 次(含 r 次)的概率为

$$P\{r \leqslant m \leqslant n\} = \sum_{m=r}^{n} C_n^m p^m (1-p)^{(n-m)}$$

(2)成功的次数不大于 r 次(含 r 次)的概率为

$$P\{0 \leqslant m \leqslant r\} = \sum_{m=0}^{r} C_n^m p^m (1-p)^{(n-m)}$$

(3)成功的次数在 s 与 r 次之间($r > s$,含 r 次和 s 次)的概率为

$$P\{s \leqslant m \leqslant r\} = \sum_{m=s}^{r} C_n^m p^m (1-p)^{(n-m)}$$

可以用 Excel 中的 BINOMDIST 函数计算随机变量 ξ 取某个数值的概率,或取值自 0 到 m 的累积概率或积分概率。该函数赋值如下 BINOMDIST(成功次数 m,试验次数 n,单次试验成功的概率 p,开关值)。当开关值赋值“false”,返回 $\xi = m$ 的概率;赋值“true”时,返回 $\xi \leqslant m$ 的累积概率。例如输入 BINOMDIST(2,4,0.5,FALSE),是计算 $p = 0.5$ 条件下 4 次实验成功 2 次的概率,返回 0.375。如果输入 BINOMDIST(2,4,0.5,TRUE),则计算 $p = 0.5$ 条件下 4 次实验成功次数不大于 2 的概率,即试验 4 次,成功 2 次、1 次和 0 次的概率之和,返回 0.6875。

4.4.2 二项式分布的性质

(1) 二项式分布是离散型数值变量的分布,当试验次数为 n 时,变量有$(n+1)$个取值,分别为$(0,1,2,\cdots\cdots,n)$。

(2) n 次试验,变量取值为 m 的概率为 $C_n^m p^m (1-p)^{(n-m)}$。

(3) 总的概率和为 1,即 $\sum_{m=0}^{n} C_n^m p^m q^{(n-m)} = 1$。

(4) 二项式分布的数学期望值 $E(\xi) = np$。

数学期望值的定义是

$$E(\xi) = \sum_{m=0}^{n} m C_n^m p^m q^{(n-m)} = np \tag{4-19}$$

这里不去证明公式(4-19)的成立。数学期望可以理解为总体的平均值,n 次试验平均成功的次数为 np 次,这应该是很自然,可以理解的。

(5) 二项式分布的方差 $D(\xi)$ 定义为

$$D(\xi) = \sum_{m=0}^{n} (m-np)^2 C_n^m p^m q^{(n-m)} \tag{4-20}$$

根据公式(3-6)可以计算得到

$$D(\xi) = E(\xi^2) - (E(\xi))^2 = npq \tag{4-21}$$

(6) 当 $p = q = 0.5$ 时,数学期望等于$0.5n$,而且分布是对称的,即成功 m 次和失败 m 次的概率相等。

4.4.3 二项式分布的应用实例

例题一 出10道是非题测验学生的水平。如果某学生没有认真学习,随意“瞎答”,问他答对6题以上的概率多大。测试数 $n = 10$,因为是随意“瞎答”,是非题答对和答错的可能都是一半,即 $p = q = 0.5$。计算 $P\{6 \leqslant r \leqslant 10\} = \sum_{m=6}^{10} C_{10}^{m} p^{m} q^{(10-m)} = 37.7\%$。这可以用 Excel 中的 BINOMDIST 函数计算。先算答对0到5题的累计概率 $P\{r \leqslant 5\}$ = BINOMDIST(5,10,0.5,TRUE),返回0.623。因此答对6题或6题以上的概率,即“瞎答”及格的概率为(1 - 0.623) = 37.7%。如果出20道题,“瞎答”答对12题以上及格的概率,应该是[1 - BINOMDIST(11,20,0.5,TRUE)] = [1 - 0.748] = 25.2%。可见增加是非题的数目,可以降低考试结果的随机性,更真实地反映学生的水平。考古学研究中鉴定墓地成年人骨的性别,判断成年男女性比是否正常,在方法上与这类是非题考试的情况十分类似的,为了提高判别性比情况的可信度,鉴定的人骨数目必须足够多,我们在以后的章节中将详细讨论。

例题二 前面的例题中,$p = q = 0.5$。这里我们分析一个 $p \neq q$ 的更普遍的情况。还是出10道题测验,但是为选择题,从5个答案中选一个正确的答案。这样“瞎答”正确的概率 $p = 0.2$,“瞎答”错误的概率 $q = 0.8$。计算“瞎答”及格的概率 $P\{6 \leqslant r \leqslant 10\} = \sum_{m=6}^{10} C_{10}^{m} p^{m} q^{(10-m)}$ = [1 - BINOMDIST(5,10,0.2,TRUE)] = [1 - 0.9936] = 0.64%。对于10道从5个答案中选一个正确答案的选择题,“瞎答”及格的概率小于1%,显然比10道是非题能更真实地反映学生的实际水平。

现在计算机普及了,Excel 等软件使得二项式分布的计算变得方便简单。而在计算机普遍应用前,当 n 和 m 数值很大时,二项式计算十分繁琐复杂。所幸当 n 和 m 数值很大时($n \geqslant 30, m \geqslant 5$),二项式分布趋向于正态分布,可以通过正态分布来处理,从而显著地简化了计算过程,详细情况将在第八章中介绍。

4.5 正态分布

上节讨论的二项式分布适用于离散型的随机变量,变量取值局限于正整数范围。更多的随机变量是可以取值小数和分数的,例如人体的身高,器物的尺寸,聚落的面积,青铜器和陶器中化学元素的含量等,它们称为连续型随机变量。连续型随机变量一般用正态分布来处理。正态分布又名高斯分布,是著名的德国数学家高斯在研究误差理论时首先提出的。正态分布函数是概率统计学和误差理论中最重要的分布函数,很多其他的分布函数,如 t 分布函数等都是根据正态分布函数扩展推导出来的。因此正态分布无愧于被称为分布之母。此外在现实世界中很多变量取值的经验分布也是服从、或者接近于正态分布的。例如某个种族成年男性的身高,某地区某时段生产的陶瓷器中各元素的含量,本书前面章节中关于 Dorset 地区巨石文化石柱柱洞的直径(见图3-2)和后面要介绍的山东半岛蛤堆顶贝丘遗址第3层贝壳的宽度(见图4-4)等变量的分布都十分接近正态分

布。而且以后我们还将看到,即使某变量本身的分布偏离正态分布,该变量平均值的分布也必定会趋向于正态分布。例如前面4.3.2中也提到,当样本容量很大时,二项式分布趋向于正态分布。但为了更好地了解正态分布需要先介绍关于概率密度、概率密度函数和定积分三个基本概念。

4.5.1 关于频率密度、频率密度函数和定积分的基本概念

我们还是从Dorset 35个石柱洞直径的例子入手引入分布的频率密度概念。柱洞直径数据自小至大整理如下:25, 27, 28, 30, 34, 35, 38, 38, 38, 39, 40, 40, 40, 42, 43, 43, 43, 44, 45, 47, 47, 47, 48, 48, 48, 48, 48, 49, 50, 50, 53, 57, 57, 58, 66厘米。第三章的图3-2a和图3-2b分别表示10厘米和5厘米区间的频率分布的直方图,可以看到这两张图的纵坐标高度是不一样的。图a把直径范围等分为5个区间,每区间10厘米,区间宽,则相应每区间的频率值高,最高的频率值为51.4%;图b把直径范围等分为9个区间,每区间5厘米,区间窄,则相应每区间的频率值就低,最高仅为28.6%。如果将每段的纵坐标频率值被分段的宽度去除,得到的商值就是单位区间宽度的频率值,这个"商值"称频率密度。表4-1显示了对Dorset 35个石柱柱洞直径数据的整理,以及计算频率分布和频率密度分布的过程和结果。图4-3a和4-3b显示了这些柱洞直径数据以10厘米和5厘米间隔的频率密度分布。对比这两张图和表4-1的左右面都可以见到,对应于相同直径值的频率密度值是相当接近的,例如两图上频率密度的峰值分别为5.14和5.72(%/cm)。这个现象反映了频率密度与区间宽度基本无关。当然在这两张图上和表4-1的左右面,对应于相同直径值的频率密度值之间还是有一些差别的,特别是在直径值的高低两端。这是因为柱洞的总数太少(才35个)、每个直径区间中所包含的柱洞数量更少,柱洞数量的随机涨落导致了频率密度差别的存在。因此不能把直径区间分划得太窄,以免每个区间中的柱洞数目太少。

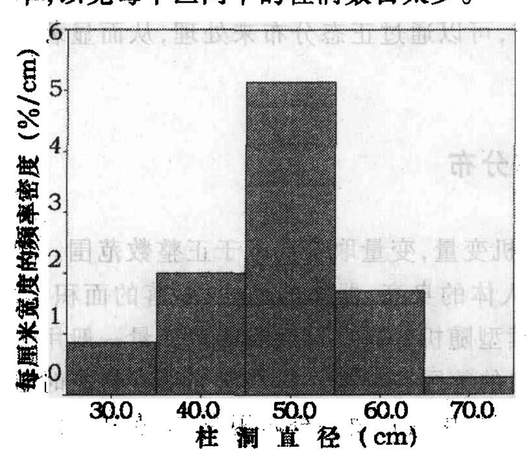


图4-3a Dorset地区巨石文化石柱柱洞直径测量值的频率密度分布直方图(a)以10cm为间隔

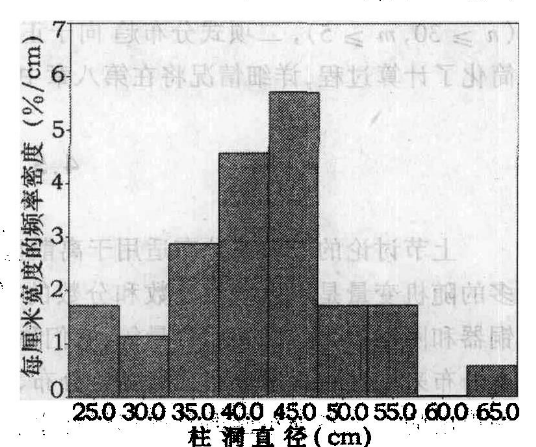


图4-3b Dorset地区巨石文化石柱柱洞直径测量值的频率密度分布直方图(b)以5cm为间隔

表 4-1 Dorset 地区巨石文化柱洞直径测量值按频次、频率和频率密度的分布表

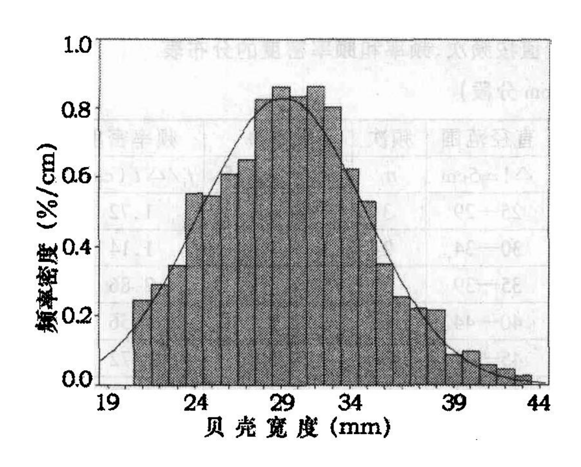
(以 10cm 和 5cm 分段)

直径范围	频次	频率%	频率密度	直径范围	频次	频率%	频率密度
△l = 10cm	n_i	$f_i = n_i/35$	$f_i/\triangle l$ (cm^{-1})	△l = 5cm	n_i	$f_i = n_i/35$	$f_i/\triangle l$ (cm^{-1})
20—29	3	8.57	0.86	25—29	3	8.6	1.72
				30—34	2	5.7	1.14
30—39	7	20.00	2.0	35—39	5	14.3	2.86
				40—44	8	22.8	4.56
40—49	18	51.43	5.14	45—49	10	28.6	5.72
				50—54	3	8.6	1.72
50—59	6	17.14	1.71	55—59	3	8.6	1.72
				60—64	0	0	0
60—69	1	2.86	0.29	65—69	1	2.9	0.58
总和	35	100.00				100.00	

当所研究样本中实体的数目很多时,可以把属性的取值范围分成很多组,每组的区间很窄,但每个区间中仍可有相当数量的实体。这种情况下每个区间所对应的频率密度也就趋向稳定。袁靖(Yuan,2003)在山东半岛牟平县蛤堆顶贝丘遗址的第 3 层,采集并测量了 2300 多个 Venerupis variegata 种贝壳的宽度。图 4-4a 是约 2300 个贝壳个体按其宽度的频率密度分布直方图(宽度小于 20 毫米的贝壳数均合并在 20 毫米的贝壳中,因此图中 20 毫米处显示的频率密度比实际情况要偏高)。因为样本容量甚大,对于从 20 毫米到 43 毫米的贝壳宽度,可以每 1 毫米分一组,共分成 24 组,而每组的个体数仍有几十到接近 200。我们把图上每个直方形上端的中心点用平滑的曲线连起来,得到图 4-4b,直方图趋向了平滑的曲线。这条曲线称为频率密度曲线,可以看出它与图 4-4a 中拟合直方图的正态曲线相当接近。虽然在图 4-4a 中,每组的个体数和相应每组的频率密度值都还有一定的涨落,因而曲线 4-4b 也还有一些小的起伏。但从图 4-4b 的频率密度分布曲线可清楚看出,贝壳宽度的中心值在 29 毫米左右,短于 20 毫米和宽于 42 毫米的贝壳极少。因为 20-42 毫米已包含了绝对大多数贝壳的尺寸的范围,图 4-4a 中全部长方形面积之和,以及图 4-4b 频率密度曲线下面的面积都应该等于 1(或十分接近于 1)的。利用这条频率密度曲线,对于蛤堆顶遗址第 3 层的贝壳可以方便地计算,宽度处于某两个具体数值 a 和 b 之间的贝壳数占有多大比例。为此从横轴的 a 和 b 两点作横轴的垂线,它们和频率密度曲线相交于 a′和 b′两点,图形 aa′bb′所组成的面积就给出这个比例值。

当实体的数量不断增加时,实体的分组可以更细,而且根据本章第一节关于概率的定义(公式 4-1),频率趋向于概率,同时频率密度曲线上那些小的起伏和涨落也逐步平滑消失,频率密度曲线趋向概率密度曲线。如果把属性的取值作为自变量,用 x 表示,那么反映概率密度曲线的函数就用 $f(x)$ 表示。这时计算 x 取值在 a, b 之间的概率($P\{a \leqslant x \leqslant b\}$),要用数学上所谓的定积分的方法:

$$P\{a \leqslant x \leqslant b\} = \int_a^b (x)\,dx \tag{4-22}$$



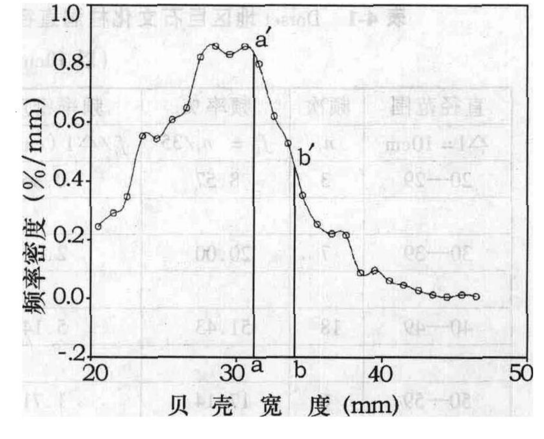


图 4-4a 蛤堆顶贝壳宽度的频率密度分布(a)
频率密度分布直方图及正态拟合曲线

图 4-4b 蛤堆顶贝壳宽度的频率密度分布(b)
频率密度分布曲线

公式(4-22)中的"∫"是积分符号,它是一个拉长了的"S","S"是英文字"Sum"(总和)的第一个字母。积分在某种意义上就是求和。a 和 b 称为定积分的上下限,表明需要计算概率密度曲线 $f(x)$ 下面由区间 $[a,\ b]$ 界定的面积。上面的式子读为"对函数 $f(x)$ 从 a 到 b 的定积分",它是一个确定的数值。需要说明,对于离散型的随机变量,则可以讨论该变量取某个数值的概率;而对于连续型的随机变量,讨论它取某个数值的概率是没有意义的,应该讨论它取值在某个区间 $[a,\ b]$ 的概率。对于离散型的随机变量,需要了解它的概率分布函数,而对于连续型的随机变量,需要了解它的概率密度函数 $f(x)$。有了上面的基本知识,可以讨论正态分布函数。

4.5.2 正态分布函数及其性质

正态分布的函数是概率密度函数,它的分析形式是

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4-23}$$

式中的 π 和 e 分别是圆周率和自然对数底两个常数。正态函数包含两个参数 μ 和 σ,因此它也常被写成 $N(\mu,\sigma)$。后面可以看到 μ 和 σ^2 分别是正态函数的数学期望值(总体平均值)和方差。图 4-5 显示了正态函数的分布图。

从公式(4-23)和图 4-5 可以看到,正态分布有下列的性质:

(1) 它形似"钟形",以 $x = \mu$ 为中心,中间高两端低且左右对称,即有

$$f(\mu - x) = f(\mu + x) \tag{4-24}$$

在 $x = \mu$ 处,$f(x)$ 取极大值,等于 $\left(\frac{1}{\sqrt{2\pi}\sigma}\right)$,曲线向左右延伸,$f(x)$ 取值不断变小,趋向于零而不等于零,正态曲线以横轴为渐近线。

(2) 曲线下面的面积与 μ 和 σ 无关,总是等于 1。或者说正态分布函数从 $-\infty$ 到 ∞ 的积分值等于 1。因为随机变量在 $[-\infty, \infty]$ 区间取值是一个必然事件。

$$P\{-\infty \leqslant x \leqslant \infty\} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = 1 \tag{4-25}$$



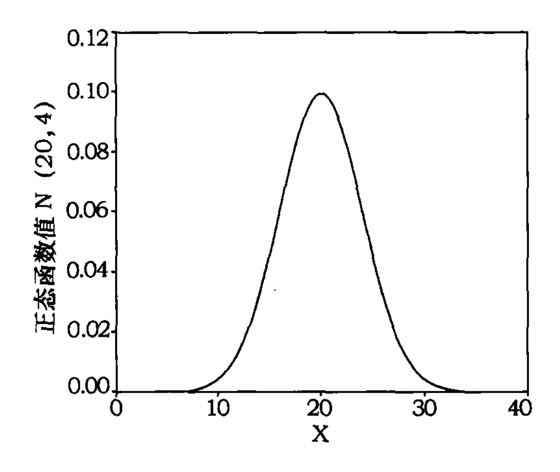


图 4-5　正态分布函数图($\mu=20,\sigma=4$)

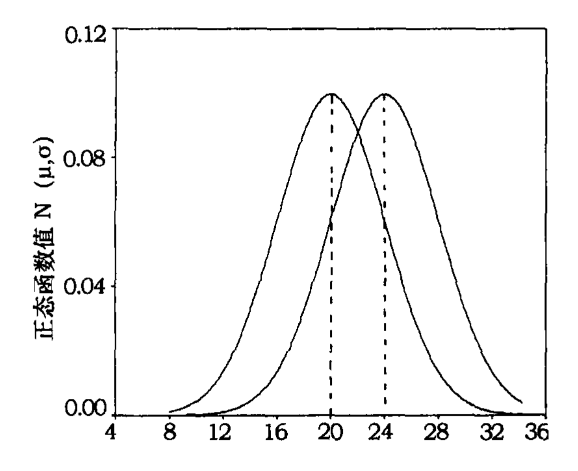


图 4-6　μ 值变化($\mu=20$ 和 $\mu=24$)导致正态分布曲线左右移动,但曲线的形状不发生变化

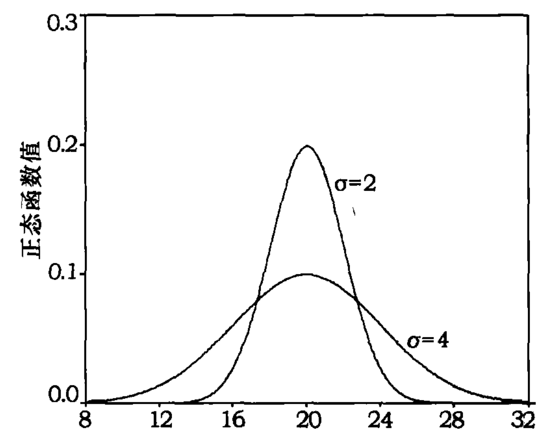


图 4-7　σ 值变化($\sigma=2$ 和 $\sigma=4$)导致正态分布曲线峰值高矮和曲线胖瘦的变化,但曲线的中心位置不变

公式(4-25)称为正态函数的归一化条件。

(3) 当 μ 变化时,图形左右移动而形状不发生变化,如图 4-6 所示;当 σ 变大时,曲线位置不动但峰值变低图形变“胖”,反之,当 σ 变小时,峰值变高图形变“瘦”,如图 4-7 所示。

(4) 可以证明 μ 和σ^2 分别是正态分布的数学期望值 $E(x)$ 和方差 $D(x)$。在 3.2 节讨论求分组样本的平均值时,表 3-3 曾经讨论,用各组的中数与频率乘积的累计和 $\left(\sum_{i=1}^{n} f_i M_i\right)$ 来替代样本的总平均值,而且当样本容量 n 很大时,两者趋向一致。因此类似有下面的式子求 $E(x)$

$$E(x) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu \tag{4-26}$$

本书不可能去求解这个定积分的数值,只是写出结果为 μ。同理可计算方差 $D(x)$

$$D(x) = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2 \qquad (4\text{-}27)$$

4.5.3 标准型正态分布

正态函数包含 μ 和 σ 两个参数，实际应用时计算比较麻烦。需要将正态函数公式(4-23)转换成对不同的 μ 和 σ 都适用的标准型正态函数公式。为此对变量 x 做如下变换：

$$Z = \frac{x - \mu}{\sigma} \qquad (4\text{-}28)$$

Z 称为 Z 分量或称标准分。将 Z 的表达式(4-28)代入公式(4-23)，可以得到

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}} \qquad (4\text{-}29)$$

公式(4-29)是标准型的正态分布函数。对于标准型的正态分布，它的数学期望值 $E(Z) = 0$，和方差 $D(Z) = 1$。公式(4-28)所执行的变量转换的功能是：(1)把正态曲线平移，使其中心移到坐标原点位置；(2)改变横轴的度量尺度，使得 $\sigma = 1$，即用标准差 σ 作为横坐标的度量单位。对于 Z 同样可以写出：

(1) 归一化条件

$$P\{-\infty \leqslant Z \leqslant \infty\} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}} dZ = 1 \qquad (4\text{-}30)$$

(2) 数学期望值 $E(Z) = 0$ (4-31)

(3) 方差 $D(x) = \sigma^2 = 1$ (4-32)

标准型正态分布函数常写为 $N(0,1)$，它实际上是一般型正态分布 $N(\mu,\sigma)$ 的一个特殊形式。两者之间很容易相互转换，在后面介绍正态分布的应用实例中经常要进行转换。下面是根据公式(4-28)得到的一张转换表。

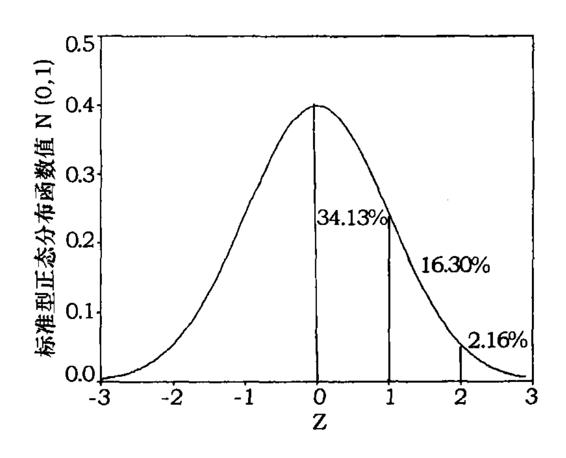
表 4-2 x 与 Z 间的转换关系

x	μ	$\mu+\sigma$	$\mu-\sigma$	$\mu+2\sigma$	$\mu-2\sigma$
Z	0	1	-1	2	-2

图 4-8 和表 4-3 显示了标准型正态分布曲线下各部分的面积，也就是 Z 在一定取值范围内的概率。

表 4-3 正态分布的 Z 和 x 在一定取值范围内的概率

Z 的取值范围	相应 x 的取值范围	概率值(%)
$-0.674 \leqslant Z \leqslant 0.674$	$\mu - 0.674\sigma \leqslant x \leqslant \mu + 0.674\sigma$	50
$-1 \leqslant Z \leqslant 1$	$\mu - \sigma \leqslant x \leqslant \mu + \sigma$	68.3
$-2 \leqslant Z \leqslant 2$	$\mu - 2\sigma \leqslant x \leqslant \mu + 2\sigma$	95.5
$-3 \leqslant Z \leqslant 3$	$\mu - 3\sigma \leqslant x \leqslant \mu + 3\sigma$	99.7
$-\infty < Z \leqslant 1$	$-\infty < x \leqslant \mu + \sigma$	84.13
$Z \geqslant 1$	$x \geqslant \mu + \sigma$	15.87
$-\infty < Z \leqslant -1$	$-\infty < x \leqslant \mu - \sigma$	15.87



4-8　标准型正态分布图

表 4-3 中所列是一些常用数值，最好能记熟在心，方便于经常使用，表的最后 2 行反映了正态函数分布的左右对称性。

在任何一本统计学的书中都可查到标准正态函数表。但表中一般不给出标准型正态函数的概率密度值 $f(Z)$，而是列出 Z 小于某个数值的累积概率值 $\Phi(Z)$。

$$P\{-\infty \leqslant \xi \leqslant Z\} = \Phi(Z) = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} d\xi \tag{4-33}$$

标准正态函数表也可以反过来查，即已知某累积概率值 $\Phi(Z)$，查 Z 应等于什么数值。

使用微软的 EXCEL 软件，可以方便地得到正态分布函数的有关数值。NORMDIST (x,μ,σ, true)返回小于某个数值的累积概率 $\Phi(x)$，例如输入 $x=\mu+\sigma$，返回的应是 0.8413。而 NORMDIST (x,μ,σ, false)，返回的是赋值为 x 时正态函数本身的数值，即正态函数的概率密度值，正态函数的概率密度值在一般的统计学书中是不易查到的。例如输入 NORMDIST(0, 0, 1, false)，返回的应是 $\frac{1}{\sqrt{2\pi}}=0.399$，即标准型正态函数的峰值。EXCEL 软件还给出正态函数反函数的数值，输入累积概率值，返回相应的 Z 值。函数形式是 NORMINV(累积概率值，μ,σ)。例如输入 NORMINV(0.8413，μ,σ)，返回的是($\mu+\sigma$)，输入 NORMINV(0.5，0，1)，返回的是 0。

4.5.4　正态分布的应用实例

(1) 美国调查统计了五年级学生的智商 IQ 值，表明 IQ 值服从正态分布，IQ 的平均值为 100，标准差为 15。现在要问某个五年级学生的 IQ 需高于多少，才能列入最聪明的 10%之中。

解：已知 $\Phi(Z)=1-0.1=0.9$

查标准正态函数表(反查)，得到 $Z=1.28$

计算(IQ－100)/15 ＝1.28，解此方程得到 IQ＝119.2

答案：五年级学生的 IQ 需高于 119.2，才能列入顶尖的 10%。

(2) 假设某地区某年有 40000 人报考大学理工科，录取 20000 人。已知每个考生的总分接近正态分布，平均分为 545 分，标准差为 30 分，问有多少人因 1 分之差而未被录取。

因为录取人数为报考人数的一半,考生的平均分也就是录取分数线。

解:查 EXCEL 的 NORMDIST 函数,

NORMDIST (x, μ, σ, true) = NORMDIST(544, 545, 30, true) = 0.4867

已知“545”是平均分,根据正态函数的对称性必然有 NORMDIST(545, 545, 30, true) = 0.5000

差 1 分未被录取的人数 = 总报考人数 × 考分在[544,545]区间的概率

40000 × (0.5 − 0.4867) = 533(人)。

答案:4 万考生中有 533 人以 1 分之差而未被录取。

(3) 碳十四测年所给出的数据也是服从正态分布的。譬如说碳十四实验室报告对某个样品的测年结果是“公元前 2460 ± 40 年”,其中 2460 年是实际测量结果,而 40 年是测量的标准差。因为服从正态分布,这个报告的含义是,样品的碳十四年龄有 68.3% 的概率落在公元前[2500,2420]的年代区间,有 95.5% 的概率落在公元前[2540,2380]的年代区间。

(4) 本节前面我们曾提到牟平县蛤堆顶贝丘遗址第 3 层 2300 多个 Venerupis variegata 种贝壳宽度的测量值接近于服从正态分布。如果这个推论正确,那么应该有大约 68.3% 贝壳的宽度值处于 $[\mu-\sigma,\mu+\sigma]$ 区间中。希望验证实际上宽度处于 $[\mu-\sigma,\mu+\sigma]$ 区间的贝壳的百分数是否接近理论值 68.3%。μ 和 σ 是蛤堆顶遗址第 3 层这类贝壳全部个体的宽度的平均值和标准差,是未知的。对这 2300 多个个体所组成的样本,其个体宽度的平均值和标准差是可以根据测量值计算的,已知 $\overline{X}=28.72$ (mm)和 $s=5.11$(mm)。因为样本的容量足够大,可以用 $\overline{X}$ 和 s 作为 μ 和 σ 的估计量,$[\mu-\sigma,\mu+\sigma]$ 的范围就是[33.83, 23.61](mm)。统计得到,宽度在此区间的贝壳数为 1550 个,除以总个体数 2329,计算得到宽度落在此区间的贝壳数占总数的 66.3 %,与正态分布的理论值 68.3% 相当接近。也可以统计宽度小于 $\mu+2\sigma=28.72+2\times5.11=38.94$ (mm)的贝壳数,为 2270 个,占总数的 97.3 %,与理论值 $\Phi(2)$ = NORMDIST(2,0,1,TRUE) = 0.9773 也十分接近。上面的比较结果可作为该类贝壳的宽度服从正态分布的验证。

第五章 统计推断和总体参数的估计

5.1 考古总体和考古样本,统计推断的基本思想

考古学是根据实物遗存资料去复原古代社会的科学。但是考古发掘的资料相对于古代社会来说总是零星不完整的资料,两者间是局部(或称样本)与全局(或称总体)的关系。另外还需考虑到,古代实物遗存长期埋在地下会受到人为和自然的破坏,遗存的被发现在一定程度上是随机的。所以英国过程主义考古学的创始人 D.L.Clarke 在关于考古学的定义中为“实物遗存资料”加了修饰词,说考古学是根据“零星不完整”,而且往往是“被扭曲了”的实物遗存资料去复原古代社会的一门科学。

举例来说我们想知道某地区某时段聚落的平均面积(μ)有多大,理论上应该测量该地该时所有聚落的面积,然后计算它们的平均值。但这是不可能实现的,很多古代聚落今天已不再保存,或者还没有被发现。我们可能调查测量了 n 个这类聚落,可以计算由这 n 个聚落面积所组成的样本的平均面积 $\overline{X}$。$\overline{X}$ 是根据已发现的诸聚落面积的实际观测值 X_i 计算得到的,因此 $\overline{X}$ 与 X_i 一样都是随机变量。很自然地我们会考虑用样本的平均值 $\overline{X}$,去估计该地区该时段所有聚落的平均面积 μ,或者说用 $\overline{X}$ 作为 μ 的估计量。

再举一个例子。黄蕴平(1996)曾对周口店第一地点和南京汤山两地肿骨鹿下颌骨 M_3 处的平均厚度 μ_1 和 μ_2 作测量比较,观测它们之间的差异有多大,以作为判断两个动物群在时代上是否能区别早晚的旁证。具体的做法是测量两地实际发现的若干个肿骨鹿下颌骨 M_3 处的厚度,再计算它们的平均厚度 $\overline{X}_1$ 和 $\overline{X}_2$,并用两个样本的平均厚度的差别去估计两个动物群中肿骨鹿下颌骨 M_3 处的平均厚度 μ_1 和 μ_2 之间的差别。

第三个例子。周礼《考工记》记录,“金有六齐,……,三分其金而锡居一,谓之大刃之齐”。是记录战国时期冶铸青铜剑的合金配方中,锡含量的设计值应为$\frac{1}{3+1}=25\%$。为了判断这个记录是否正确,可以通过测量一批 n 把现存的战国青铜剑的锡含量,比较它们的平均值与“25%”差别有多大,来判断六齐说关于青铜剑的合金配方是否符合实际。

上面三个例子都是涉及总体和样本间的关系。我们看到总体是被研究对象的全体,即全部应研究实体的集合;而样本是从总体中按一定方法抽取出来的有限数目(n 个)实体的组合。前两个例子是用样本的平均值或平均值之差去推断去估计总体的平均值或平均值之差,属于对总体参数的估计。第三个例子是使用样本的平均值去判断关于总体平均值的某个理论假设是否符合实际,属于关于总体参数的假设检验。总体的参数估计和关于总体假设的检验是统计推断的两个主要方面。

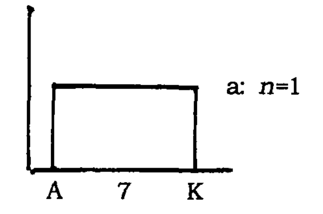
可以看出,对考古资料的研究分析,是用有限的考古资料去推断古代社会情况,是一

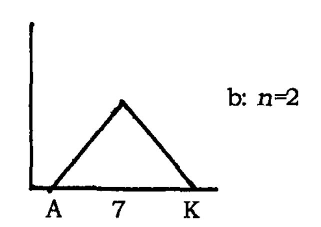
个由样本推断总体的统计推断过程，所得的结论只具有统计学的意义，而不应看作绝对真理。以这些结论作为前提进行逻辑推理所获得的考古学新的知识同样是带有统计性质的。统计推断有其本身的特点和规则，将是本章和后面几章讨论的内容。本章主要介绍总体参数的估计，而第六、七章讨论假设检验。

5.2 样本平均值的分布和样本的标准误

5.1 中的三个例子显示了统计推断的两个方面：用样本的平均值估计总体的平均值和用样本的平均值去检验有关总体参数的假设。这两方面的问题都与样本的平均值有关。因此首先要讨论样本平均值的有关性质。

为了便于理解，将通过日常生活中的例子来进行讨论。假设已知中国成年男子的平均身高为 172 厘米，标准差为 5 厘米，而且身高 X_i 接近于正态分布。因为身高服从正态分布，约有 68.3% 的成年男子其身高在 167 到 177 厘米之间。如果现在随机地抽取 n 个个体（譬如说 100 人），并测量了他们的身高 X_i，可以计算出这个样本的平均值 $\overline{X}_1$。再随机另外抽取 100 人，又可以得到第二个样本的平均值 $\overline{X}_2$。继续抽取个体，可以得到多个（譬如说 r 个）容量为 $n = 100$ 的样本。每个样本有一个平均值 $\overline{X}_j(j = 1\cdots r)$。这些 $\overline{X}_j$ 相互间一般是不相等的，它们是一个新的随机变量。我们需要研究这个新随机变量的分布，它的数学期望值、方差和标准差。





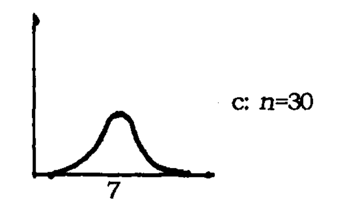


图 5-1 回放抽取扑克牌的均匀分布情况下，样本容量 $n = 1$，$n = 2$ 和 $n = 30$ 时样本平均值的概率分布图（引自 Spate，1989）

5.2.1 样本平均值 $\overline{X}$ 的分布

男性身高的原始分布是接近正态分布的，而一般情况下随机变量的分布可能是各种各样的，例如 4.2 中介绍的均匀分布和二项式分布等。在实际中还可能出现双峰的分布、不对称的分布等。在统计学中有一条著名且重要的定理，叫中心极限定理。它能证明，不论原始的分布是什么形式，只要样本的容量 n 足够大（一般定 $n \geqslant 30$），样本平均值 $\overline{X}$ 的分布总是接近正态分布的，n 愈大，分布愈趋近正态。本书不可能来证明这个定理，而是通过从均匀分布总体中抽取的样本，样本平均值随样本容量增加而趋向正态分布的例子，来显示这个"趋向"过程。图 5-1a 是随机抽一张扑克牌得到的 X 取值从 1 到 13 的均匀分布（本图和图 4-1 是同一图），其总体平均值 $\mu = 7$。如果抽了一张牌后放回再抽第二张，即随机并独立抽取 2 张扑克牌，其平均值的概率分布如图 5 - 1b 所示。分布已从均匀分布的长方形变成了左右对称的三角形，分布的中心

仍在"7"处。分布的"趋中"可以这样理解,分布的最外两端例如2张牌的平均值为13或1这类事件的出现,必须要求2次抽样结果都是13或者都是1,这类事件的概率是1/169。而2张牌平均值为7的事件,可以通过2次抽7,第1次6和第2次8,第1次8和第2次6,……等共15种过程来实现,因此先后独立抽取2张牌,平均值为7的概率为15/169。平均值的分布范围仍为1—13,但中央部位出现的概率要高于两端的,而且相对于中心是左右对称的。还需指出,单次抽牌只可能有1—13共13种结果;而2张牌的平均值除13个整数外,还会出现1.5—12.5等12个半整数。随着抽样次数的增加,平均值虽仍介于1—13间,但可能取值的数值却不断增加,并愈益趋向连续化。图5-1c是30次随机独立抽扑克牌,即$n=30$的样本的平均值的概率分布。从图看出它已非常接近正态,而且这个新随机变量的总体平均值仍是"7",且左右对称。这张图还显示,随n的增大,样本平均值分布趋向正态的过程是很快的。用统计学的术语是,收敛很快。

5.2.2　样本平均值$\bar{X}$的数学期望和方差

原始观测数据X_i是从一个其平均值为μ,方差为σ^2的总体中随机抽取出来的,即有$E(X_i)=\mu$和$D(X_i)=\sigma^2$。可以求容量为n的样本的平均值$\bar{X}$的数学期望和方差

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\left(\sum_{i=1}^{n}E(X_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \tag{5-1}$$

$$D(\bar{X}) = D\left(\sum_{i=1}^{n}X_i/n\right) = \frac{1}{n^2}\sum_{i=1}^{n}D(X_i) = \frac{\sigma^2}{n} \tag{5-2}$$

这两个公式显示样本平均值$\bar{X}$的数学期望和单次观测值X_i的数学期望值是相等的,都是μ。而样本平均值$\bar{X}$的方差却比单次观测值X_i的方差小,前者是后者的n分之一,为$\frac{\sigma^2}{n}$。

因为总体的方差σ^2往往是未知的,经常用样本的方差s^2来取代,这样对于样本平均值的方差可以写出下面的公式

$$s_{\bar{X}}^2 = \frac{s^2}{n} \tag{5-3}$$

而样本平均值$\bar{X}$的标准差为

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \tag{5-4}$$

$s_{\bar{X}}$又称为样本的标准误。标准误是原始观测数据的标准差s的$\sqrt{n}$分之一。在5.3.1小节中我们将说明样本方差s^2是总体方差σ^2的最佳估计量。

综合5.2.1和5.2.2两小节的内容,结论如下:不论原始观测数据X_i服从什么分布,只要样本的容量足够大$n \geqslant 30$,样本的平均值$\bar{X}$服从以总体的平均值μ为数学期望值,以总体方差σ^2的n分之一为方差的正态分布。

回到我国成年男子身高的例子,100个男子身高的平均值$\bar{X}$服从正态分布,$\bar{X}$和单个男子身高X_i有相同的数学期望值$\mu = 172cm$,但前者的方差小得多,仅为后者的一百分之

一，或者说前者的标准差仅为后者的十之一。已知 X_i 的标准差为 5cm，前面已提到，约有 68.3%的成年男子个体其身高在 167 到 177 厘米之间。现在随机抽取了很多组男子，每组都是 100 人，那么组平均身高的标准差为 $\frac{5}{\sqrt{100}} = 0.5$cm，即约有 68.3%的组其平均身高处于 171.5 到 172.5 厘米之间。显然各组平均身高之间的涨落远小于个体之间身高的涨落。

5.3 总体方差的点估计和大样本总体平均值的区间估计

需要指出，本节下面讨论仅限于 $n > 30$ 的大样本的情况。总体平均值 μ 和总体方差 σ^2 是总体的两个参数，它们有固定的数值，但经常是未知的。而样本平均值 $\overline{X}$ 和样本方差 s^2 是随机变量，是由实际观测结果计算而得的。经常需要用实测数据来对总体参数作估计，或者说把后者作为前者的估计量。如果用一个确定的数值去估计总体参数，称为总体参数的点估计，但经常用一个数值范围去估计总体参数，称为总体参数的区间估计。

因为对总体平均值的区间估计涉及对总体方差的估计，因此先讨论总体方差 σ^2 的点估计。

5.3.1 总体方差 σ^2 的点估计

第三章公式(3-4)给出了方差的计算公式 $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 但是这样计算的样本方差 S^2 并不是总体方差 σ^2 的最佳估计。而计算样本方差的公式(3-7)，即 $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 才给出总体方差 σ^2 的最佳估计。所谓最佳估计是要求估计量满足无偏、有效和一致性等三个条件。为什么分母上要用 $(n-1)$ 替代 n 才能得到 σ^2 的最佳估计呢，这里不可能对此作详细的讨论。我们仅指出计算样本方差的公式本来应该是 $D(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$，但在公式(3-4)计算 S^2 时用 $\overline{X}$ 替代了未知的 μ，从而使得 S^2 的数值偏小，小于 $D(X)$。为此把公式(3-4)的分母 n 相应改成 $(n-1)$，以作补偿。同样的理由需要用样本的标准差 s，而不是 S 作为总体标准差 σ 的估计量（严格地说，s 并不是 σ 的最佳估计，但是对于 σ 的估计，s 是优于 S 的）。

关于总体方差 σ^2 的区间估计将在 5.5 中讨论。

5.3.2 总体平均值 μ 的点估计和区间估计

样本平均值 $\overline{X}$ 和任何一个实际测量值 X_i 原则上都可以作为总体平均值 μ 的估计量。但在前一节已看到 $\overline{X}$ 比 X_i 的离散性小，有更大的概率接近于总体的平均值 μ，显然作为 μ 的估计量 $\overline{X}$ 优于 X_i（见图 5-2）。当然也可以用样本的中数，甚至样本的几何平均值来估计 μ，但在所有的估计量中 $\overline{X}$ 是 μ 的最佳点估计。另一方面，当平均值的概念应用于连续

性的数量属性时,不太适宜用一个确定的数值作点估计,而更适宜于用一个数值区间来估计,后者称为总体平均值的区间估计。

这里自然会想到用$[\bar{X}-s_{\bar{X}},\bar{X}+s_{\bar{X}}]$作为总体平均值$\mu$的区间估计。这个区间以$\bar{X}$为中心,2倍的标准误$s_{\bar{X}}$为宽度。下面以Dorset地区巨石文化柱洞直径的数据为例讨论μ的区间估计。这个样本测量了$n=35$个数据,已知样本的平均值为$\bar{X}=43.80$ cm,标准差为$s=9.03$ cm。计算得到这个样本的标准误$s_{\bar{X}}=\frac{9.03}{\sqrt{35}}=1.53$ cm,因此相应的估计区间为43.80 ± 1.53 cm,即$[42.27, 45.33]$ cm。现在必然要提出的问题是,这个估计区间的置信度有多高,即有多大的可能性μ落在这个区间之中。因为在本例中,样本的容量为$35>30$,属大样本,$\bar{X}$应接近于正态分布。上述区间$[\bar{X}\pm s_{\bar{X}}]$的宽度为2个标准误,可知这个区间估计的置信度应为68.3%,即有68.3%的可能性总体平均值μ处于这个区间之中。$[\bar{X}\pm s_{\bar{X}}]$称为置信度为68.3%的置信区间。对此也可以理解成:如果我们有100有同样容量的样本,那么100个$[\bar{X}\pm s_{\bar{X}}]$区间中,大致有68个区间把总体平均值μ包含其中。区间$[\bar{X}\pm s_{\bar{X}}]$也可以写成$\left[\bar{X}\pm\frac{s}{\sqrt{n}}\right]$或$\left[\bar{X}\pm\frac{\sigma}{\sqrt{n}}\right]$,因为$s$是$\sigma$的估计量。

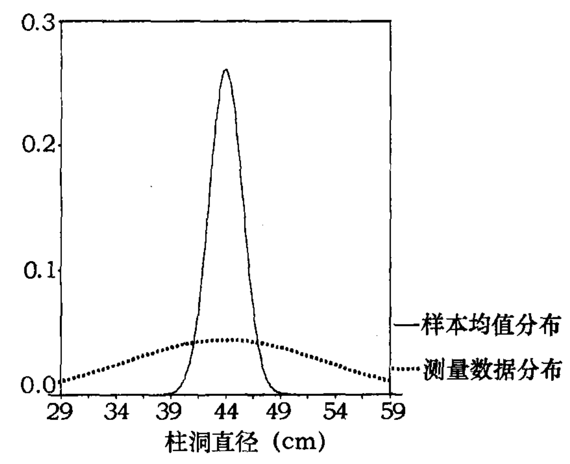


图5-2　以Dorset地区柱洞直径样本为例,显示样本平均值比单次测量值接近总体平均值的概率更高,后者应为43.8cm左右

上面的讨论是给定估计区间的宽度后,求估计的置信度。反过来怎样在设定置信度要求的条件下寻找相应的估计区间呢。定义$\alpha=(1-\text{置信度})$,α称为显著性水平。譬如说希望找置信度为95%,即显著性水平$\alpha=1-0.95=0.05$的区间估计。因为置信区间是以$\bar{X}$为中心向两侧伸展的,而正态分布曲线以μ为中心左右对称的,应该找正态分布函数的累积概率函数$\Phi(Z)=\frac{\alpha}{2}$的位置$Z_{\frac{\alpha}{2}}$。置信度为$(1-\alpha)$的估计区间应该是

$$[\bar{X} \pm Z_{\frac{\alpha}{2}} s_{\bar{X}}] \text{或} \left[\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right] \tag{5-5}$$

当 $\alpha = 0.05$ 时，找查 $\Phi(Z) = 0.025$ 的位置 $Z_{0.025}$，可以通过查正态函数表，或用 EXCEL 软件 NORMINV(0.975,0,1)函数找 $Z_{0.025}$ 的值，$Z_{0.025} = 1.96$。因此相对于显著性水平为 0.05 或置信度为95% 的置信区间是$[\bar{X} \pm 1.96 s_{\bar{X}}]$或$\left[\bar{X} \pm 1.96 \frac{s}{\sqrt{n}}\right]$。对于 Dorset 地区纪念性建筑物柱洞直径的例子，已知 $\bar{X} = 43.80\text{cm}$，$s_{\bar{X}} = 1.53\text{cm}$，计算得 $1.96 \times 1.53 = 3.00\text{cm}$，由此显著性水平为 0.05 的总体平均值的置信区间是[43.80 ± 3.00]cm 或[40.80,46.80]cm。

5.3.3 总体平均值区间估计中置信度、置信区间宽度和样品容量三者间的关系

从公式(5-5) 可以看到，用样本的平均值估计总体平均值时，(1) 显著性水平 α 或置信度$(1-\alpha)$、(2) 置信区间的宽度$\left(2Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$和(3) 样本的容量 n 等三个量之间存在确定的关系。在样本容量 n 一定的条件下，显著性水平 α 值越小，置信度越高，置信区间也越宽。置信区间的宽度反映估计的精密度。因此估计的置信度和精密度之间是互相制约的。提高其中的一个指标是以牺牲另一个指标为条件的。为了在不变的置信度下，提高估计的精密度，唯一的方法是增大样本的容量，获取更多的观测数据。当然增加观测数据是以付出更多的研究精力和经费为条件的，而且对于考古研究而言，增加观测数据有时客观上是不被允许的。

再次回到 Dorset 地区巨石文化柱洞直径的例子，上面已计算得到这个 $n = 35$ 的样本，其平均值为 43.80cm，标准误 $s_{\bar{X}} = \frac{9.03}{\sqrt{35}} = 1.53\text{cm}$。置信度为 68.3% 的估计区间[42.27,45.33]cm 的宽度为 3.06cm。如果希望提高置信度到 95%，置信区间就应放宽到 $2 \times 1.96 \times 1.53 = 6.00\text{cm}$，即用区间[43.80 ± 3.00]cm 去估计。如果希望仍在 95%置信度下，置信区间的宽度减为 3cm，就应该增加测量数据。至少应测量多少个柱洞的直径呢？列出方程

$$2 \times 1.96 \times \frac{9.03}{\sqrt{n}} = 3$$

解此方程，得 $n = 139$，需要测量 139 个柱洞的直径。在同等的置信度的条件下，要使置信区间缩窄到原来宽度的一半，即估计的精密度提高 1 倍，必须把观测的数据量增加 4 倍。当然对于 Dorset 地区柱洞直径的例子，有可能找不到这么多的柱洞。

5.4 观测数据少的小样本的总体平均值的估计和 t 分布

5.4.1 t 分布函数及其性质

上节讨论了大样本情况下对总体平均值的估计。但是有时候所掌握的观测数据量

很少，n < 30。这种情况在考古学研究中，特别是在旧石器考古研究中经常会遇到的。鉴于人类活动的遗存长年埋于地下而受到破坏和丢失，考古学家总是苦于材料的贫乏。当然对于这类观测数据量很少的小样本同样可以计算它们的样本平均值 $\bar{X}$，不过小样本的平均值 $\bar{X}$ 一般不服从正态分布。因此上节讨论的内容不能照搬应用于小样本，需要引进一个新的、小样本平均值所服从的分布函数，称为 t 分布函数。统计学中证明如果小样本所来自的总体服从正态分布，即单个测量值 X_i 服从正态分布，那么小样本的平均值 $\bar{X}$ 服从 t 分布。或者更正确地说，对小样本平均值 $\bar{X}$ 标准化后所得的统计量：

$$\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$$

服从自由度 $df = (n-1)$ 的 t 分布。也可写为

$$\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t(n-1) \tag{5-6}$$

这里引进了统计学中广泛运用的关于自由度的概念，自由度在这里可以简单地理解为观测数据的数目 (n) 减去为决定某些量所使用的关系的数目。在定义 t 的公式(5-6) 中用了样本的标准差 s，而计算 s 时，默认和使用了 $\sum(X_i-\bar{X}) = 0$ 这个关系式，因此样本的 n 个元素中只有$(n-1)$个是独立的，自由度等于$(n-1)$。t 分布的函数形式较复杂，我们仅显示不同自由度 t 分布函数的图，如图 5-3。

从图可见 t 分布函数的形状与标准型的正态分布很相似，只是概率密度曲线总体上被压低、拉宽了。t 分布有如下的一些性质，其中(1)至(4)可以从图 5-3 直接看出：

(1) 与标准型的正态分布，即 Z 分布类似，t 分布以 $t = 0$ 为中心，左右对称，随 t 的绝对值增加，函数迅速下降，趋近于零。

(2) t 分布函数的值总是正的，而曲线下的总面积等于 1。

(3) t 分布比 Z 分布的离散性大，即峰值偏低而分布宽。t 分布的自由度 df 越高，离散性越小，当 df 大于 30 时，t 分布将十分接近于正态分布，当 df 趋向无限大时，t 分布将趋向正态分布。

(4) t 分布的数学期望 $E(t) = 0$。

(5) t 分布的方差 $D(t) = \frac{df}{df-2}$，当 df 很大时，t 分布的方差趋近于 1。

在任何一本统计学书中都附有 t 函数表。可以查表得到不同自由度的 t 函数的累积概率值。也就是说，给定显著性水平 α，可以查到各个自由度情况下的 t_α 值，使得 $P\{t \geqslant t_\alpha\} = \alpha$ 。但受限于书的版面，t 函数表只能对有限的若干个显著性水平值 α 列出相应的 t_α 值。较为方便的是使用 Excel 软件的 TDIST 和 TINV 两个函数找 t 函数值，可以查到任何 α 值和任何自由度时的 t 值或相应的累积概率。TDIST (t, df, 双侧或单侧) 函数是根据已知的 t 值和自由度，求双侧或单侧的累积概率。按序输入 t 值，自由度值，和开关值“2”或者“1”，函数相应返回双侧累积概率 $P\{|t| > x\}$ 或单侧累概率 $P\{t > x\}$。例如输入

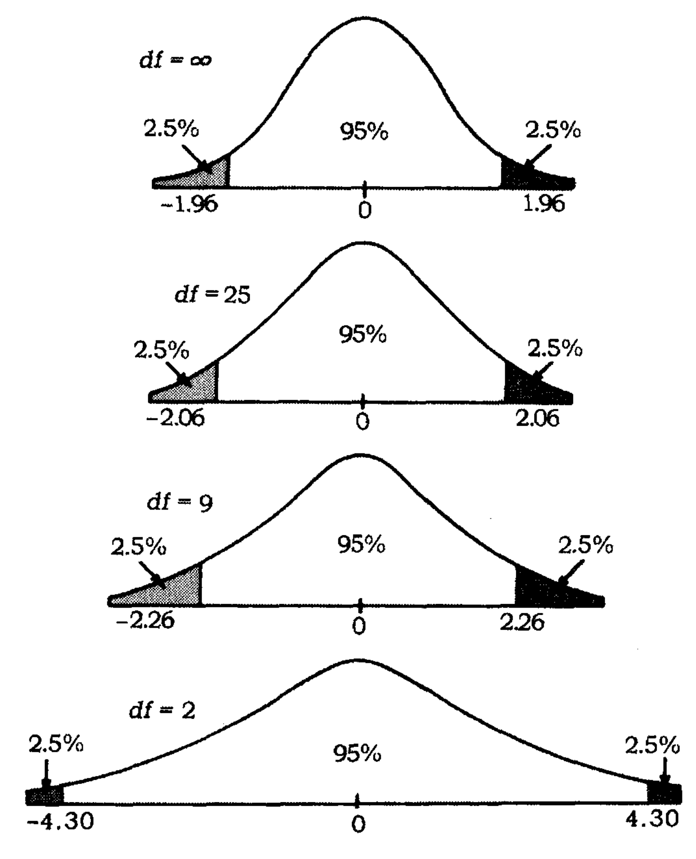


图 5-3　不同自由度的 t 分布函数图(引自 Spatz,1989)

TDIST(2,30,2)，返回 $df = 30$ 条件时，$P\{t > 2\} + P\{t < -2\} = 0.0546$；输入 TDIST(2,30,1)，返回 $P\{t > 2\} = 0.0273$。

TINV(双侧累积概率，自由度)是 TDIST 的反函数，按序输入双侧累积概率和自由度，返回相应的 t 值，使得 $P\{t > |x|\}$ = 输入的双侧累积概率值。例如输入 TINV(0.05,30)，返回 2.04，使得 $P\{t > 2.04\} + P\{t < -2.04\} = 0.05$。

有了上面关于 t 分布函数的基本了解，可以返回讨论小样本情况下，根据样本的 n，$\bar{X}$ 和 s 对总体平均值 μ 的区间估计。给定显著性水平 α，查表可找到 $t_{\frac{\alpha}{2}}$，区间 $\left[\bar{X} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$ 就是置信度为 $(1-\alpha)$ 对 μ 的区间估计。

5.4.2　小样本总体平均值的区间估计

在某旧石器遗址的一个地层中发掘出 13 片石片，它们的重量分别为 14.3，14.1，13.6，13.5，12.0，11.5，11.3，10.9，10.6，9.8，9.7，9.3，7.8（克）。希望用这个样本，以 90% 的置信度来估计该地层中石片的平均重量 μ。

计算样本的平均重量 $\bar{X} = 11.42$(克)

样本的标准差 $s = \sqrt{\dfrac{\sum_{i=1}^{13}(X_i - \overline{X})^2}{13-1}} = 2.02$(克)

样本平均值 $\overline{X}$ 的标准差 $s_{\overline{X}} = \dfrac{2.02}{\sqrt{13}} = 0.56$(克)

自由度 $df = 12$

取置信度$(1-\alpha) = 90\%$，$\dfrac{a}{2} = 0.05$

查表或计算 TINV(0.10,12) = 1.78，$1.78 \times 0.56 = 1.10$(克)

对 μ 的 90%置信度的估计区间[10.32,12.52](克)

要求不同的置信度，会得到不同宽度的估计区间。

最后需要补充说明一点。本节开始时，曾要求样本所来自的总体应该服从正态分布，这种条件下样本的平均值 $\overline{X}$ 才服从 t 分布。在考古研究中，经常不清楚原始观测数据是否服从正态分布，还因为观测数据的数量少，无法画出它们的经验分布图来检验观测数据是否与正态分布接近。因此难以判断使用 t 分布的假设前提是否成立。所幸 t 分布的宽容度相当大，即使总体的分布相当程度地偏离于正态分布，从中抽样所得样本的平均值的分布仍接近于 t 分布。因此在绝大多数情况下，处理考古数据的平均值时我们是可以用 t 分布的。

5.5　χ^2 分布函数和总体方差的区间估计

5.5.1　样本方差的分布和 χ^2 分布函数

在5.3中曾提到样本方差 s^2 是总体方差 σ^2 的最佳估计，和用样本的标准差 s 作为总体标准差 σ 的估计量。s^2 是根据诸实测数据 X_i 计算而得的，它也是一个随机变量，也有其分布的规律。一般情况下 s^2 的分布比较复杂，但是如果已知 X_i 来自正态总体 $N(\mu,\sigma^2)$，那么可以证明，统计量

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} = \frac{s^2(n-1)}{\sigma^2} = \chi^2(n-1) \tag{5-7}$$

服从自由度为$(n-1)$的 χ^2(读作卡方)分布 $\chi^2(n-1)$。

χ^2 实际上是以 σ^2 为度量尺度的样本的离差平方和。χ^2 函数是一个单参数的函数，唯一的参数是自由度 df。χ^2 函数的分析表达式比较复杂，图 5-4 是几个不同自由度的 χ^2 分布图

χ^2 分布图与 Z 分布和 t 分布不一样，它不是左右对称的。但 χ^2 分布曲线下的面积也是等于 1 的。可以证明 χ^2 分布的数学期望等于其自由度 df

$$E(\chi^2(df)) = df \tag{5-8}$$

而 χ^2 分布的方差等于 2 倍的自由度

$$D(\chi^2(df)) = 2(df) \tag{5-9}$$

χ^2 分布函数的值可以查表，也可以用 Excel 软件中的 CHIDIST 和 CHIINV 两个函数来

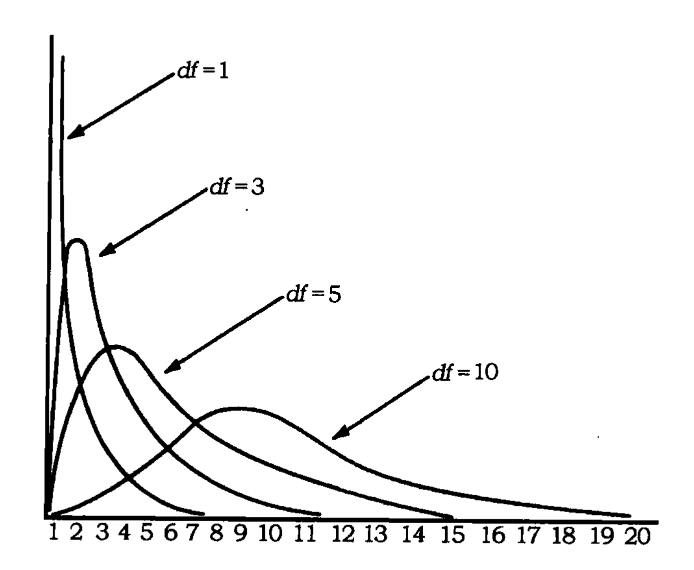


图 5-4 不同自由度的 χ^2 分布函数图(图引自 Spatz(1989))

计算。因为 χ^2 分布的非对称性,表中总是给出单侧的累积概率数值。CHIDIST 函数的形式是 CHIDIST(x,自由度),x 是用来计算分布的值,返回的是 $\chi^2 > x$ 的概率 $P\{\chi^2 > x\}$。例如要计算自由度 = 12,$\chi^2 > 20$ 的概率,在 Excel 文件中键入 CHISIST(20,12),将返回 $P\{\chi^2 > 20\} = 0.067$。反函数 CHIINV(单侧累积概率数值,自由度)返回对应的 χ^2 值。例如自由度 $df = 12$ 时,要找 x,使得 $P\{\chi^2 > x\} = 0.067$,键入 CHIINV(0.067,12),将返回 $x = 20$。

5.5.2 总体方差 σ^2 的区间估计*

本章3.1中已介绍样本的方差 s^2 是总体方差 σ^2 的最佳点估计,并且知道了在总体服从正态分布的条件下,s^2 服从自由度为 $(n-1)$ 的 χ^2 分布。现在可以在一定的显著性水平 α 下给出 σ^2 的置信区间。由于 χ^2 分布不对称,区间的上下限 $(\chi^2_{\frac{\alpha}{2}})_L$ 和 $(\chi^2_{\frac{\alpha}{2}})_H$ 需要分别找。下面还是通过 Dorset 地区巨石文化石柱洞直径测量数据的例子加以说明。已知测量了 $n = 35$ 个数据,样本的标准差 $s = 9.03\text{cm}$,s^2 应该是 $81.54\ \text{cm}^2$。求 $\alpha = 0.05$ 时总体方差 σ^2 的置信区间。计算函数

$$(\chi^2_{\frac{\alpha}{2}})_L = \text{CHIINV}(0.025,34) = 51.966$$

$$(\chi^2_{\frac{\alpha}{2}})_H = \text{CHIINV}(0.975,34) = 19.806$$

利用式(5-7)

$$\sigma_L^2 = \frac{(n-1)s^2}{\chi^2_{(\frac{\alpha}{2})}} = 34 \times \frac{81.54}{51.966} = 53.36\ \text{cm}^2$$

$$\sigma_H^2 = \frac{(n-1)s^2}{\chi^2_{(1-\frac{\alpha}{2})}} = 34 \times \frac{81.54}{19.806} = 139.98\ \text{cm}^2$$

$$\sigma_L = 7.31\,cm \qquad \sigma_H = 11.83\ \text{cm}$$

区间[53.36, 139.98]cm^2是总体方差σ^2的95%置信度的区间估计，而[7.31, 11.83]cm是总体标准差σ的95 %置信度的区间估计。注意标准差估计区间的中心位置是9.67cm，与σ的最佳估计值$S=9.03$cm是不重合的，即最佳点估计值并不处于估计区间的中央。为了在不变的显著性水平下提高对总体方差和标准差估计的精密度，唯一的方法同样是增加观测的数量，也是以增加研究经费和延长研究时间为代价的。

第六章　大样本条件下总体平均值的假设检验

第五章讨论了怎样根据样本的平均值和方差对样本所属总体的平均值作估计。本章将讨论统计推断的另一个重要方面——假设检验，主要是关于总体平均值的假设检验。

考古研究中经常要求比较两个同层次实体的某个数值属性之间有没有差别。例如比较两个地区同时期的遗址密度是否有差异，以探讨古人对居住地环境的选择是否有倾向性。或者观察某种动物的某种形态特征的测量值在前后两期间是否发生了变化，以探讨它的进化。考古研究中还可能碰到的另一类问题是，检验实际观测的考古资料是否符合某种理论模式。例如《六齐说》记录东周时期铸造青铜剑的配方，其锡的含量应为25%。现测量了一批东周青铜剑的锡含量，要根据实测的青铜剑的锡平均含量去检验《六齐说》的配方是否符合实际。在回答上述问题作判断时，我们所依据的是，或者比较两个样本的平均值 $\bar{X}_1$ 和 $\bar{X}_2$，或者是用样本的平均值 $\bar{X}$ 与理论值 μ 作比较。样本的平均值是随机变量，因而是有涨落的。常见的情况是 $\bar{X}_1$ 和 $\bar{X}_2$ 并不精确相等，$\bar{X}$ 与 μ 也可能不绝对相等。必须要确定一个数值标准，当 $\bar{X}_1$ 与 $\bar{X}_2$ 差别，或者 $\bar{X}$ 与 μ 的差别要达到多大，才可以认为这个差别超出了随机涨落的范围，从而判断所研究的两个总体的平均值(μ_1, μ_2)之间，或者样本的数学期望值与理论模式的平均值 μ 间确实存在差别。假设检验的目的就是要寻求这样一个判断标准。同时需要指出，在假设检验中，无论是作出肯定或者否定的判断，都有一定的可能性判断错误。假设检验的过程在作判断时，应该能同时给出判断错误的概率大小。因此假设检验所作的推断是带有统计性的，不是“绝对真理”。

20 世纪 60 年代，过程主义考古学派十分强调假设检验在考古研究中的地位。他们对古代社会、对古人的行为和活动模式提出各种假设，然后用实际的考古资料，甚至设计新的考古发掘来验证所提出的假设是否成立。他们认为，能通过各种检验的假设是最符合实际情况的假设，是最强的假设，甚至可上升为关于古代社会的理论。

最后应说明，本章的讨论仅局限于大样本情况下总体平均值的假设检验。

6.1　大样本单总体 U 检验的原理和实例

U 检验是指利用正态分布进行的关于总体平均值的假设检验，单总体是指检验单个样本的数学期望值是否与总体平均值一致。本节将通过检验《六齐说》大刃之齐关于青铜剑中锡含量的配方是否为 25%，和检验碳十四测年数据与所测墓墓主人死亡年代的关系等两个实例，来说明 U 检验的基本过程。

6.1.1　大刃之齐锡含量的 U 检验

《考工记》记有“金有六齐，……，三分其金而锡居一，谓之大刃之齐”。就是说在东周

时期,青铜剑冶铸的合金配方是锡含量占25%(对“六齐”的另一种解释,认为青铜剑配方的锡含量占33%,我们取低值25%)。华觉民(1999)统计并发表了43把东周青铜剑铜锡铅的百分组成(数据引自《中国古代青铜技术》的表7-11)。这43把剑锡含量的平均值为 $\overline{X} = 16.27\%$,标准差 $s = 2.42\%$(下面的书写省略单位“%”)。现在要检验这43把剑中锡含量的数学期望值与理论值 $\mu = 25$ 之间是否有显著的差别,或者说这43把剑是否来自一个其锡的百分含量平均值为25的总体,是否按照锡含量为25%的配方铸造的。具体的检验过程分成四步。

第一步,假设这43把青铜剑所组成的样本的数学期望值与理论值一致,称为原假设,写作 H_0: $E(X) = \mu = 25$。

相应有一个备择假设 H_1: $E(X) \neq \mu = 25$。

第二步,寻找一个可用以检验的统计量。显然这个统计量应该是与样本的平均值有关的。因为青铜剑的数量为43把,属大样本,其锡含量的平均值 $\overline{X}$ 应服从标准差为 $\frac{s}{\sqrt{n}}$ 的正态分布。如果原假设 H_0 成立,那么由公式(6-1)定义的 Z 应该服从标准型的正态分布。所以选择 Z 作为检验用的统计量。

$$Z = \frac{|\overline{X} - \mu|}{\frac{s}{\sqrt{n}}} \tag{6-1}$$

计算Z的数值

$$Z = \frac{|\overline{X} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|16.27 - 25|}{\frac{2.42}{\sqrt{43}}} = \frac{8.63}{0.369} = 23.66$$

式(6-1)的分子是样本平均值 $\overline{X}$ 与理论值 μ 的差值。由于要检验的仅是否存在差别,并不理会哪个量的大或小,所以取绝对值。这称为双侧或双尾的检验。式(6-1)的分母是样本的标准误。

第三步,选择检验的显著性水平 α。显著性水平决定了接受或拒绝 H_0 的置信度,一般取 α 等于0.10, 0.05或0.01等数值。这里我们取 $\alpha = 0.05$,利用正态分布函数表可以查到对应的检出阈,或称为判别域 $Z_{\frac{\alpha}{2}} = 1.96$,$Z_{\frac{\alpha}{2}}$ 是确定接受或拒绝 H_0 时 Z 的取值范围。因为是双侧的检验,所以检出阈是 $Z_{\frac{\alpha}{2}}$,而不是 Z_α。

第四步,作判断。根据计算得到的统计量 Z 与检出限 $Z_{\frac{\alpha}{2}}$ 的大小作判断。如果 $Z < Z_{\frac{\alpha}{2}}$,那么接受原假 H_0;反之,如果 $Z > Z_{\frac{\alpha}{2}}$,则拒绝原假 H_0 而接受备择假设 H_1。这样判断的依据是:在 H_0 成立的条件下,根据正态分布出现 $Z > Z_{\frac{\alpha}{2}}$ 的概率是 α,而 α 是一个很小的值(本例中选择了 $\alpha = 0.05$)。统计学中有一条“小概率原理”,认为在单次试验中小概率事件是不可能出现的,如果这种小概率事件竟然出现了,我们就应怀疑 H_0 的合理性,从而拒绝 H_0。本例中 $Z = 23.66 > Z_{\frac{\alpha}{2}} = 1.96$。因此根据实测的43把青铜剑锡的百分含量,在 $\alpha = 0.05$ 的显著性水平上,拒绝了《考工记》记录大刃之齐锡含量为25%的说法。也可以在 H_0 成立的条件下,计算出现 $Z \geqslant 23.66$ 的概率 α,在这个例子中 $\alpha < 10^{-30}$,如此小

的概率事件在单次试验中是不可能出现的，但实际情况是这种事件出现了，从而只能拒绝原假设 H_0。

上面的判断等同于确定一个区间 $\left[(\bar{X}-\mu)-Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}},(\bar{X}-\mu)+Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$，观察它是否把坐标轴的原点包含在内，或者说这个区间的上下限是否是一个取正值，另一个取负值。如果答案是肯定的，那么接受 H_0，反之则拒绝 H_0，接受 H_1。对于本例，这个区间是[8.73 ± 1.96 × 0.369]，或[8.00, 9.45]，当然拒绝 H_0，接受 H_1。这个判断过程也可以理解为，用样本的标准误 $\frac{s}{\sqrt{n}}$ 作为标准尺度，去衡量实测的样本平均值 $\bar{X}$ 和理论值 μ 差别 $|\bar{X}-\mu|$ 的大小，当这个差别小于标准误 $\frac{s}{\sqrt{n}}$ 的 $Z_{\frac{\alpha}{2}}$ 倍时，就认为差别更可能是随机过程引起的，因而接受原假设。如差别 $|\bar{X}-\mu|$ 大于 $Z_{\frac{\alpha}{2}}$ 倍的标准误时，则拒绝原假设 H_0。

从上面的讨论可以看到，接受或拒绝原假设，取决于(1) $\bar{X}$ 与 μ 之间的差距，(2) 样本的标准误 $\frac{s}{\sqrt{n}}$ 和(3) 选择的显著性水平 α。

6.1.2 用东周青铜剑的锡铅含量之和检验大刃之齐

6.1.1 节在讨论东周青铜剑的锡含量的假设检验的四步过程中，掺杂了对方法的说明和解释。为了把假设检验的过程阐述更为简明清晰，现用青铜剑的锡铅含量之和替代单一的锡含量，再进行四步的假设检验。古人在撰写《考工记》时，未知能否分辨锡和铅是两种不同的金属，即《大刃之齐》中的“三分其金而锡居一”，锡是否可能包含了锡和铅两种金属之和。学术界多数认为在战国时，甚至更早，古人已能分辨锡和铅。我们对此并不质疑，这里把锡和铅的含量合在一起作为六齐说中的锡来对待，只是作为假设检验的一个例子。统计上述 43 把青铜剑中锡铅含量和的平均值及其标准差为(20.40 ± 4.44)%。现检验如下：

(1) 原假设 $H_0:E(X)=\mu=25$；　　备择假设 $H_1:E(X)\neq\mu$。

(2) 选检验用统计量并计算其数值：

$$Z=\frac{|\bar{X}-\mu|}{\frac{s}{\sqrt{n}}}=\frac{|20.44-25|}{\frac{4.44}{\sqrt{43}}}=\frac{4.56}{0.677}=6.73。$$

(3) 选择确定显著性水平 $\alpha=0.01$，查 $Z_{\frac{\alpha}{2}}=Z_{0.005}=2.58$。

(4) 因为 $Z=6.73>Z_{0.005}=2.58$，在 $\alpha=0.01$ 水平上，拒绝 H_0，接受 H_1。

检验的结论是：根据 43 把东周青铜剑化学组成的实际测量结果，在 $\alpha=0.01$ 的显著性水平上，不能认为东周青铜剑是依据一个锡铅含量和为 25 % 的配方铸成的。也就是说，即使把锡和铅加在一起，青铜剑中它们的含量和也达不到“六齐”所要求的 25%。关于“在 $\alpha=0.01$ 的显著性水平上”这个限定词的含义是，至少有 99% 的把握拒绝原假设，后面我们还将对此作详细讨论。既然拒绝了“东周青铜剑是用一个锡铅含量和为 25% 的配方铸成的”这样一个原假设，就可以对“东周青铜剑实际锡铅含量和的数学期望值”和

“‘六齐’所要求的25%”之间的差别作区间估计。在 $\alpha = 0.01$ 的显著性水平上，这个区间是 $4.56 \pm 2.58 \times 0.677 = (4.56 \pm 1.75)\%$。看来《大刃之齐》“三分其金而锡居一”的记录离实际情况甚远。

6.1.3　碳十四测年结果的 U 检验

发现了一座墓葬，取墓主人的人骨做碳十四测年，结果为公元 300 ± 80 年。从各方面的材料分析，这很可能是某一位王的墓，而且历史记载，该王死于公元 200 年。现在检验碳十四测年结果与该王已知的去世年代间有无显著差异。

已知碳十四年龄是服从正态分布的，而“80 年”是测年过程给出的标准差。因此可以进行 U 检验，检验过程如下：

(1) H_0　碳十四年龄与某王的死亡年龄间无显著差异，$X = \mu$；

(2) 计算统计量 Z：$Z = \frac{|X - \mu|}{\sigma} = \frac{300 - 200}{80} = 1.25$；

(3) 确定显著性水平 $\alpha = 0.05$，查正态函数表得判别阈 $Z_{0.025} = 1.96$；

(4) 因为 $Z = 1.25 < Z_{0.025} = 1.96$，在 $\alpha = 0.05$ 的显著性水平上保留 H_0，即认为该墓的测定年代与历史记载中某王的去世年代没有矛盾。保留原假设并不是说测年结果证明了该墓墓主人的死亡年代就是 AD 200 年，而只是表明测年结果与 AD200 年间没有显著的差别。

再次强调，上面三个假设检验的例子中所作出的推论都是统计性质的，无论是接受或拒绝原假设，都有一定的可能性犯错误。关于假设检验中的错误问题，将在 6.3 节中进行详细讨论。

6.2　双侧检验和单测检验

6.1 节的三个实例均是检验样本的数学期望值和某个理论值或已知值之间有没有差别，并不在乎它们间谁大谁小，因此属于双侧的假设检验。但有的情况下要回答的问题不仅是是否有差别，而且要了解差别的方向，明白谁大谁小。例如，已知解放前我国北方男子的平均身高是 170cm，现随机抽查了 200 名北方男子的身高，计算得到这个样本的平均值为 171.1cm，标准差为 6cm。问解放后由于生活水平的改善，北方男子的平均身高增高了吗？这是属于单侧的假设检验。相对于双侧的检验，单侧检验中的备择假设 H_1 和怎样确定判别域与双侧检验有所不同。现对上面的例题作检验如下：

(1) 原假设 $H_0: E(\overline{X}) = \mu$；　备择假设 $H_1: E(\overline{X}) > \mu$。

这里的备择假设与前面双侧检验中的备择假设是有区别的。

(2) 在 H_0 成立条件下计算统计量：

$$Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{171.1 - 170}{\frac{6}{\sqrt{200}}} = 2.59。$$

(3) 选择 $\alpha = 0.01$，查表 $Z_{0.01} = 2.33$。

与双侧检验不同之处在于，选择一定的显著性水平 α 后，单侧检验查 Z_α 值，而双侧检验查 $Z_{\frac{\alpha}{2}}$ 值。

(4) 因为 $Z = 2.59 > Z_{0.01} = 2.33$，拒绝 H_0，接受 H_1 $E(\overline{X}) > \mu$。

因为 H_1 被接受，可以对解放后我国北方成年男子平均身高的增长量作区间估计。计算增长量的点估计值 171.1 - 170 = 1.1 cm，样本的标准误 $= \frac{6}{\sqrt{200}} = 0.42$cm。选定估计的置信度为 95%，$\alpha = 0.05$，查表 $Z_{\frac{\alpha}{2}} = 1.96$。计算 $1.96 \times 0.42 = 0.82$ cm。因此解放后我国北方男子身高增长量置信度 95% 的估计区间[0.28，1.92]cm。

单侧检验在考古研究中的应用也是很广泛的，譬如检验某种动物的某个骨骼指数后期是否比前期增大或减小，钱币中的金银含量后期是否有降低，即有无贬值等，都需要作单侧的假设检验。

6.3 假设检验中的两类错误

前面介绍了几个假设检验的实例，对原假设 H_0 有接受的也有拒绝的。但由于检验的统计性质，无论是哪一种情况都可能犯错误。本节将讨论错误的种类和犯错误的概率大小。下面的表列出了在假设检验中检验正确或犯错误的四种可能情况。

	实际 H_0 真	实际 H_0 伪
判断接受 H_0	判断正确	第二类错误：纳伪
判断拒绝 H_0	第一类错误：弃真	判断正确

6.3.1 第一类错误：弃真错误

当原假设 H_0 实际上成立时，由于实际测量数据的随机涨落，使得 $Z > Z_{\frac{\alpha}{2}}$，导致我们错误地拒绝 H_0，这是弃真错误，也称为第一类错误。犯第一类错误的概率是可以预置的，当选定显著性水平为 α 时，实际上已经预置了犯第一类错误的概率不大于 α。另外也可以根据 Z 值的大小，计算犯第一类错误的概率。例如 6.1.2 节检验东周青铜剑的锡铅含量和是否按"六齐"配方时，曾计算得到 $Z = 6.73$。利用 Excel 软件的 NORMDIST 函数可以计算得到，在 H_0 为真的条件下，Z 达到 6.73 的概率小于 1.7×10^{-11}，这就是说我们拒绝 H_0 时犯弃真错误的概率小于 1.7×10^{-11}，犯弃真错误的概率极小极小，几乎为不可能。当然在 6.1.1 节检验东周青铜剑纯锡含量是否按"六齐"配方时，犯弃真错误的概率就更小了。

6.3.2 第二类错误：纳伪错误

当原假设 H_0 实际上不成立时，由于实际测量数据的随机涨落，使得 $Z < Z_{\frac{\alpha}{2}}$，导致我们错误地接受 H_0，从而犯纳伪的错误，也称第二类错误。纳伪错误的概率经常用 β 来表示，计算犯纳伪错误的概率比计算弃真错误的概率要复杂，而且它与三个因素有关，包括 H_0 偏离实际有多大(样本的数学期望值与理论值的实际差别大小)，样本标准误的大小，以及显著性水平的选择。

下面用6.1.3节中碳十四测量"可能是某个已知王墓"年代的例子，计算纳伪错误概率的大小，并分析纳伪概率和上述三个因素间的关系。

在这个例子中已知碳十四测年的标准差为80年，要检验的是墓主人是否死于公元200年。下面计算这个例子在不同情况下犯纳伪错误的概率。如果定显著性水平 $\alpha = 0.05$，那么不管墓主人实际死亡的年代，只要实际测年结果在公元 $200 \pm 1.96 \times 80$ 年间隔中，即在区间[44,356]中，就应接受"墓主人死于公元200年"的假设，从而可能犯纳伪的错误。假设所测墓的墓主人实际上死于公元240年，计算对该墓的测年结果落在[44,356]区间的概率。

先计算上下判别阈：

$$Z_1 = \frac{44 - 240}{80} = -2.45, \qquad Z_2 = \frac{356 - 240}{80} = 1.45$$

犯纳伪错误的概率为：

$$\Phi(Z_2) - \Phi(Z_1) = \Phi(Z_2) - (1 - \Phi(-Z_1)) = 0.9265 - (1 - 0.9926) = 0.9191$$

对于墓主人死于公元240年的墓，纳伪的概率为91.9 %。

下面分别改变墓主人实际死亡年代、显著性水平和样本的标准误等三个因素，观察它们对犯纳伪错误概率的影响。

(1) 犯纳伪错误的概率与实际的偏差有关。上面计算了墓主人实际死亡年代与"被检验年代"相差40年时纳伪的概率。如果墓主人的实际死亡年代相应为公元280年和公元360年，即与"被检验年代"相差80年和160年，那么纳伪的概率有多大呢。用同样的方法计算上下判别阈 Z_1、Z_2 和犯纳伪错误的概率，相应为83%和48%。可以看到，纳伪概率的大小与实际偏离的程度是有关的，墓主人实际死亡年代与"被检验的年代值"越接近，犯纳伪错误的概率越大。

(2) 犯纳伪错误的概率与选择的显著性水平有关。依旧假设墓主人实际死于公元240年，但把显著性水平改定为 $\alpha = 0.10$，则接受区间变为 $200 \pm 1.65 \times 80$ 年，即[68,332]年。上下判别阈为 $Z_1 = \frac{68 - 240}{80} = -2.15$ 和 $Z_2 = \frac{332 - 240}{80} = 1.15$。纳伪概率 $\beta = \Phi(1.15) - (1 - \Phi(2.15)) = 0.859$。我们记得，取 $\alpha = 0.05$ 时，犯纳伪错误的概率是 $\beta = 0.919$。考虑到显著性水平的选择等同于确定犯弃真错误的概率，对应于两个 α 值得到两个不同的纳伪概率 β 值，当犯弃真错误的概率从5% 增高到10% 时，犯纳伪错误的概率从92% 降低到86%。这表明犯第一种和第二种错误的概率是相互牵制的，试图降低犯一种错误的概率，必然以增加犯另一种错误的概率为代价。在测量误差和测量次数固定的条件下，不能要求同时降低弃真和纳伪的概率。

(3) 犯纳伪错误的概率与标准误的大小有关。仍假设墓主人实际死于公元前240年，原假设和显著性水平 $\alpha = 0.05$ 的条件也不变。但是碳十四测年重复测量了4次，4次重复测量导致测年结果的标准误差减小一半，达40年。这样接受区间也相应缩小为 $200 \pm 1.96 \times 40$ 年，即[122,278]年，相应的 $Z_1 = \frac{122 - 240}{40} = -1.47$ 和 $Z_2 = \frac{278 - 240}{40} = 0.47$，计算得到纳伪的概率为0.62，小于标准误差在80年时的纳伪概率0.919。增加测量的重复次数，即增大样本的容量 n，可以在弃真概率不变的条件下，降低纳伪的概率。但

是增加观测的数据量,是以增加研究工作的经费和延长研究的时间周期为代价的。特别是在考古研究中观测数据的数量是受到客观条件限制的。

在实际的假设检验中应该怎样来调配 α 和 β 的数值呢?这与实际研究的问题有关。一般说来,原假设中的 μ 是某种理论值,或者是总结了大量研究结果而归纳得出的,我们不希望原假设 $H_0: E(\overline{X}) = \mu$ 轻易地被否定。所以犯弃真错误的概率一般是选得很低的,如选择 $\alpha = 0.1$、0.05 和 0.01 等。更何况接受原假设仅仅表明,没有足够的证据去否定它,而并不是证明了它的正确和成立。犯纳伪错误概率的选择取决于实际研究课题对偏离所能容忍的程度。如果样本的数学期望与理论值 μ 有差别,但差别并不大,这时纳伪的概率 β 可能会相当大。但这也无妨大局,很小的差别本来就接近于没有差别,纳伪概率大些是可以容忍的。当然什么是"差别并不大"也是因事而论的。在工厂产品的抽样检验中,如果生产的是一般民用产品,可以容忍产品实际达到的指标与设计指标 μ 有一定的差异,即可以容忍 β 值大些,因而可以把产品检验中犯弃真错误的概率定得低些,即显著性水平定得较高(即 α 很小),以免抽样检验结果稍有偏离设计指标而将整个一批产品当作废品处理掉,造成损失。但是如果生产的是军工产品或药品,就不能容许纳伪概率 β 值很大,即使抽样检验结果反映产品实际达到的指标略有偏离设计值,也不应作为合格产品出厂。抽样检验军工产品或药品时,应该增大被检测样本的容量,同时显著性水平相应要定得低些,即 α 值不应很小。

6.4 大样本情况下两个总体平均值的一致性检验

本章前面讨论的是单个样本的数学期望值和预知值或理论值的比较,本节要讨论两个考古总体平均值的比较。在考古研究中是经常要处理两个总体平均值的比较。例如比较前后两期聚落的面积总体上有没有变化,钱币有没有贬值,两个地点动物骨骼的某种数量特征是否一致,同一墓地男女墓葬随葬品的平均数目是否相等,等等。

"大样本"要求每个样本的容量 n 都大于 30。6.3 中讨论的大样本单总体平均值假设检验的方法同样可以应用于大样本两总体平均值的一致性检验,因为每个样本的平均值 $\overline{X}_1$ 和 $\overline{X}_2$ 都服从正态分布。但是需要注意两个特殊点(1)怎样计算统计量 $(\overline{X}_1 - \overline{X}_2)$ 的标准差,和(2) 两个样本的个体间是独立的还是相关的。

6.4.1 两个独立样本间总体平均值的一致性检验:以钱币贬值等为例

独立样本区别于 6.4.2 小节将讨论的配对样本,独立样本中每个样本中实体的取值与另一样本中的实体是无关的。下面通过两个实例来讨论两个独立样本间总体平均值的比较。

例一　收集并测量一批带有两位皇帝名号的铜币的重量。统计了铜币的数量,它们的平均重量等数据并记录于下表。已知两位皇帝是前后相继的。希望分析了解在第二位皇帝即位后,铜币铸造的设计重量是否有所降低,即是否发生了贬值。

钱币分类	钱币数量 n	平均重量 $\bar{X}$(克)	标准差 s(克)
前一位皇帝	384	4.85	0.32
后一位皇帝	377	4.79	0.24

检验过程如下：

(1) 假设 $H_0:\mu_1 = \mu_2$(没有发生贬值)，备择假设 $H_1:\mu_2 < \mu_1$(发生了贬值)。

(2) 先计算统计量 $D_{\bar{X}} = (\bar{X}_1 - \bar{X}_2)$ 的标准差 。两个独立的随机变量的差值的标准差,等于它们各自标准差的平方和的开方。

$$s_{D_{\bar{X}}} = s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6-2}$$

计算如下：

$$s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{0.32^2}{384} + \frac{0.24^2}{377}} = 0.021$$

再计算统计量

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{s_{(\bar{X}_1 - \bar{X}_2)}} = \frac{4.85 - 4.79}{0.021} = \frac{0.06}{0.021} = 2.857$$

这个统计量也是服从标准型的正态分布的。

(3) 因为探讨的问题是否曾发生了铜钱币的贬值,属单侧检验。

选择 $\alpha = 0.01$，查表 $Z_{0.01} = 2.575$

(4) $Z = 2.875 > Z_{0.01} = 2.575$

拒绝原假设 $H_0:\mu_1 = \mu_2$，接受备择假设 $H_1:\mu_2 < \mu_1$。我们有99% 以上的把握认为第二位皇帝时,铜钱币的铸造是“缺斤短两”的。既然接受了备择假设 $H_1:\mu_2 < \mu_1$,就可以进一步估计铜币的重量减轻了多少,它应为$(0.06 \pm Z_{\frac{\alpha}{2}} \times 0.021)$(克),该区间估计的置信度为$(1 - \alpha)$,如果取 $\alpha = 0.05$,则贬值的区间估计为0.02—0.10克。也可以用贬值的百分率来表示,即第二位皇帝时,铜钱币铸造的设计重量降低了约$\frac{0.06}{4.85} = 1.2\%$。

与6.1节单总体平均值的假设检验相比较,这里新的内容是怎样计算两个随机变量的差值的标准差和计算统计量 Z 是用 $|\bar{X}_1 - \bar{X}_2|$ 替代 $|\bar{X} - \mu|$。

例二　袁靖等(Yuan,2002)测量统计了山东半岛北岸距今约5500年的蓬莱县大仲家贝丘遗址第3、4层,称之为 Venerupis variegate 种的贝壳的宽度。下表列出有关的描述性统计的数据：

层位	样本容量(n)	贝壳平均宽度 $\bar{X}$(mm)	标准差 s(mm)	标准误(mm)
大仲家4层	205	27.03	5.36	0.374
大仲家3层	165	31.85	4.74	0.369

第一步，假设两个地层中贝壳的平均宽度未发生变化,即提出原假设 $H_0:E(\bar{X}_1) =$

$E(\bar{X}_2)$;

第二步,计算统计量 $Z = \frac{|\bar{X}_1 - \bar{X}_2|}{s_{(\bar{X}_1 - \bar{X}_2)}}$;

为此先计算两层贝壳平均宽度差值 $(\bar{X}_1 - \bar{X}_2)$ 的标准差:

$$s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{0.374^2 + 0.369^2} = 0.525$$

$$Z = \frac{|\bar{X}_1 - \bar{X}_2|}{s_{(\bar{X}_1 - \bar{X}_2)}} = \frac{31.85 - 27.03}{0.525} = 9.18$$

在 $\mu_2 = \mu_1$ 成立的假设条件下,Z 服从标准型正态分布,从而 Z 达到 9.18 的概率几乎为零。因此可以以极大的把握作判断,两层贝壳的平均宽度是有明显差别的,第 4 层贝壳的宽度明显比第 3 层贝壳窄。在这项研究中,没有必要去估计两层贝壳的宽度的差值,这是没有多大的实际意义的。人们感兴趣为什么从第 3 层往上到第 4 层总体来说贝壳的个体变小了。两层地层的时代仅相隔二三百年,因此袁靖(Yuan,2002)等认为不是气候的变化导致的物种形态的变化,而是人类食用软体动物所造成的后果。人类的大量食用,特别是优先挑食个体大的贝壳,缩短了贝壳的期望年龄。贝壳体态的缩小是人类"使用压力"的后果。

与例一相比,例二的推论和书写形式简化了。计算了统计量 Z 后立即可以作判断,因为统计量 Z 的数值太大了。只要 $|Z| > 3.3$,就可以安全地拒绝两个样本的数学期望值相等的原假设,因为这时犯弃真错误的概率已小于 0.001 了。

6.4.2 配对实体的大样本间总体平均值的一致性检验

前面讨论了两期铜钱币的贬值,大仲家遗址两层地层中贝壳的个体尺寸的变化,所涉及的都是相互独立的随机变量。但是有的情况下,两个样本的成员之间相互是有关联的。例如为了研究两代男子身高间的变化,抽取了 n 对父子,然后比较父亲组和儿子组的平均身高。这里两组样本的容量是相等的,都是 n,而且它们的成员间是配对的。配对样本的例子在实际生活中是很多的。例如同一批样品,每个样品分成两份,分别用两种不同的方法测量某个指标,比较两种测量方法的结果之间是否存在系统的差异。在教育学研究中观察一批学生在经过某种培训后的进步,在考古研究中用不同的方法测量同一批陶片或青铜器的元素含量等,所得的观测结果都是配对样本。判断两个配对样本某个定量属性平均值是否有差别,只要 $n \geq 30$,也是用我们已熟悉的 U 检验方法,但是计算配对样本间平均值差别的标准差的公式与 6.4.1 节中独立样本的情况(公式(6-2))是不一样的。

公式(6-3)给出配对样本间平均值差别的标准差计算方法:

$$s_{D_{\bar{X}}} = s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_1^2 + s_2^2 - 2r_{12}s_1s_2}{n}} \tag{6-3}$$

与公式(6-2)相比,根号下多出了 $\frac{2r_{12}s_1s_2}{n}$ 项,其中 r_{12} 是 X_1 与 X_2 之间的皮尔逊相关

系数。第九章将介绍什么是皮尔逊相关系数,这里仅指出对于完全正相关的样本 $r_{12} = 1$;而对于独立样本 $r_{12} = 0$,公式(6-3)也就还原到独立样本的公式(6-2)。

配对样本平均值差别假设检验与6.4.1中两个独立样本平均值差别假设检验相比,除样本间平均值差别的标准差的计算方法不同,分别用公式(6-3)和(6-2)外,其他步骤是相同的。因此这里不再重复。我们将在第七章"小样本总体平均值的一致性检验"中,介绍配对样本平均值差别假设检验的具体例子。

第七章　小样本和多样本总体平均值的假设检验

第六章讨论了大样本(样本容量 $n \geqslant 30$)情况下总体平均值的假设检验，本章将讨论 $n < 30$ 的小样本的情况。在根据两个样本进行两总体平均值的一致性检验时，如果其中一个为小样本，也要作为小样本来处理。小样本的基本特点是其平均值和平均值的差，一般情况下不服从正态分布，而是服从 t 分布。进行小样本总体平均值的假设检验，需要考虑原始观测数据是否服从正态分布，原始观测数据所属总体的方差 σ^2 是否已知，两组观测数据所属总体的方差是否相等等一系列前提条件。因此比大样本情况下总体平均值的假设检验要复杂。希望读者注意这些前提条件，注意计算两个小样本的平均值差的标准差的公式。

7.1　单总体平均值的假设检验

首先要假设原始观测数据抽样自正态分布总体的样本。下面再分成两种不同的情况作讨论。

7.1.1　总体的方差 σ^2 已知

检验的目的与第六章的情况相似，需要检验样本的数学期望值 $E(X)$ 与总体平均值 μ 之间有没有显著的差异。如果总体方差 σ^2 已知，那么统计量 $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} = N(0,1)$ 依然是服从标准型的正态分布的。因此其检验步骤和第六章大样本的情况是类似的，这里不再重复论述。但是需要指出，在考古学研究中，总体的方差 σ^2 已知的样本是不多见的。

7.1.2　总体的方差 σ^2 未知

如果总体方差 σ^2 未知，那么对于小样本，统计量 $T = \dfrac{\bar{X} - \mu}{s/\sqrt{n}}$ 不服从正态分布，而服从自由度为 $(n-1)$ 的 t 分布，式中 s 为样本的标准差，n 为样本容量。下面我们以检验《六齐说》中关于"…… 六分其金而锡居一，谓之钟鼎之齐"为例，来说明小样本总体平均值的检验过程。根据《六齐说》，东周时青铜钟鼎铸造配方中锡的设计含量应为 14.3 %。华觉民(1999) 统计的东周青铜钟鼎的锡铅组成列于表 7-1。

表 7-1　东周青铜钟鼎的锡铅平均含量和标准差

器物类型	数量 n	锡平均含量及标准差 s%	锡铅和的平均含量和及标准差 s%
钟铃	11	14.72 ± 1.75	17.60 ± 4.55
鼎类容器	14	15.54 ± 2.40	20.54 ± 2.77
钟鼎合在一起	25	15.17 ± 2.14	19.24 ± 3.87

现检验钟铃的锡平均含量是否符合《六齐说》规定的 14.3%。

(1) 提出原假设，$H_0:E(\bar{X}) = \mu = 14.3, H_1:E(\bar{X}) \neq 14.3\%$。

(2) 计算统计量

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \tag{7-1}$$

$$T = \frac{14.72 - 14.3}{1.75/\sqrt{11}} = 0.782。$$

(3) 选定 $\alpha = 0.05$，利用 EXCEL 软件的 t 函数计算判断阈，$T_{0.025} = \text{TINV}(0.05,10) = 2.23$。

(4) 因为 $T = 0.782 < T_{0.025} = 2.23$，接受 $H_0:E(\bar{X}) = 14.3\%$。

检验结论为：在 0.05 的显著性水平上，东周青铜钟铃等响器的实测平均锡含量符合《六齐说》的配方。

对于由 14 件鼎等青铜容器组成的样本，可用相同的方法，计算得到 $T = 2.17$，查 $T_{0.025} = \text{TINV}(0.05,13) = 2.16$。因为 $T = 2.17 > T_{0.025} = 2.16$，所以在 0.05 的显著性水平上，认为东周青铜鼎等容器实测的锡平均含量与《六齐说》的配方有差异。如果取 $\alpha = 0.02$，则 $T_{0.01} = 3.01 > T = 2.17$，那么在 0.02 的显著性水平上，可以认为东周青铜鼎容器实测的锡平均含量与《六齐说》的配方不矛盾。这里我们再次看到假设检验的结论可能会依赖于显著性水平的选取，对于 14 件青铜容器的锡平均含量，当取 $\alpha = 0.05$ 时，拒绝《六齐说》的原假设，而取 $\alpha = 0.02$ 时，接受《六齐说》的原假设。出现这种看似矛盾的情况，是因为实测的鼎等容器的锡平均含量与《六齐说》的设计值之间的差异对应于样本的标准误差别不大所致，前者为 $(15.54 - 14.3) = 1.24$，而标准误为 $2.4/\sqrt{14} = 0.64$。

7.2　独立样本两个总体平均值一致性的假设检验

本节的讨论要求，两个样本的原始观测数据分别来自正态分布总体 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$。在本节的后面将介绍怎样粗略地验证正态分布前提是否成立。区别于总体方差 σ_1^2 和 σ_2^2 是否已知和是否相等，检验所使用的统计量也是不同的。

7.2.1　总体方差 σ_1^2 和 σ_2^2 已知

这种情况下统计量 $D_{\bar{X}} = \bar{X}_1 - \bar{X}_2$ 服从正态分布 $N\left((\mu_1 - \mu_2),\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)\right)$，式中的 n_1 和 n_2 分别为两个样本的容量。可以将 $D_{\bar{X}}$ 标准化后得到 Z，然后进行我们已熟悉的 U 检验。但是正如前面已提到，对于考古样本总体方差经常是未知的，我们不准备列举实例对此作详细的讨论。

7.2.2　总体方差 σ_1^2 和 σ_2^2 未知，但是 $\sigma_1^2 = \sigma_2^2$

如果满足 $\sigma_1^2 = \sigma_2^2$，可以用下面的公式计算它们的计权平均

$$s^2 = \frac{(n_1-1)s_1^2}{(n_1-1)(n_2-1)} + \frac{(n_2-1)s_2^2}{(n_1-1)(n_2-1)} \tag{7-2}$$

而统计量 $D_{\bar{X}} = \bar{X}_1 - \bar{X}_2$ 的方差为

$$s_{D_{\bar{X}}}^2 = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \tag{7-3}$$

将式(7-2)代入式(7-3),得到

$$S_{D_{\bar{X}}}^2 = \left[\frac{(n_1-1)s_1^2}{(n_1-1)(n_2-1)} + \frac{(n_2-1)s_2^2}{(n_1-1)(n_2-1)}\right]\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)} \times \frac{n_1+n_2}{n_1 n_2} \tag{7-4}$$

这时 $T = \dfrac{\bar{X}_1 - \bar{X}_2}{s_{D_{\bar{X}}}}$ 服从自由度为 (n_1+n_2-2) 的 t 分布。下面通过3个具体的例子来说明小样本情况下两总体平均值一致性的假设检验。这里假定小样本两总体平均值一致性 t 检验的前提条件是满足的。

实例一　已知发掘了一个墓地。墓主人性别和随葬品的数量统计如表 7-2 所示:

表 7-2　某墓地的墓主人性别和随葬品数量的统计表

墓主人性别	墓葬数	随葬品数量	平均数	标准差
男性	8	39,54,59,62,46,53,52,41	$\bar{X}_1 = 50.75$	$s_1 = 8.17$
女性	11	34,49,33,45,42,37,40,41,38,48,51	$\bar{X}_2 = 41.64$	$s_2 = 6.04$

需要检验随葬品的多寡是否与性别有关。检验过程如下:

(1) 假设随葬品的多寡与性别无关 $H_0: E(\bar{X}_1) = E(\bar{X}_2)$。

(2)计算统计量

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_{D_{\bar{X}}}} \tag{7-5}$$

为此先计算

$$s^2 = \frac{(n_1-1)s_1^2}{(n_1-1)+(n_2-1)} + \frac{(n_2-1)s_2^2}{(n_1-1)+(n_2-1)} = \frac{7\times 8.17^2 + 10\times 6.04^2}{7+10} = 48.94$$

$$s_{D_{\bar{X}}}^2 = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 48.94 \times \left(\frac{1}{8} + \frac{1}{11}\right) = 10.57$$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_{D_{\bar{X}}}} = \frac{50.75 - 41.64}{\sqrt{10.57}} = 2.80$$

(3) 选定 $\alpha = 0.02$,查 T 表 $T_{0.01}(df = 17) = 2.57$。

(4) 因为 $T = 2.80 > T_{0.01} = 2.57$,在 $\alpha = 0.02$ 水平上,拒绝原假设,认为随葬品的多寡与墓主人的性别有关,男性墓葬的随葬品平均数量多。

实例二　检验东周青铜剑和青铜戈戟的平均锡含量之差,是否与《六齐说》的设计值一致。表 7-3 列出有关数据(数据引自华觉明[1999])。

表 7-3　东周青铜剑和青铜戈戟的实测平均锡含量和《六齐说》记录的锡含量表

器物名称	数量	实测平均锡含量%	标准差%	《六齐说》的锡含量%
青铜剑	43	16.27	2.42	25
青铜戈戟	15	15.99	2.29	20

首先利用式(7-5)计算两类器物锡含量平均值差的方差 $s^2_{D\overline{X}}$：

$$s^2_{D\overline{X}} = \frac{(43-1)\times 2.42^2 + (15-1)\times 2.29^2}{(43+15-2)} \times \frac{43+15}{43\times 15} = 0.513$$

$$s_{D\overline{X}} = 0.716$$

检验过程如下：

根据《六齐说》两种青铜器锡含量设计值之差应为 25 - 20 = 5。因此：(1)提出原假设 $H_0: E(\overline{X}_1) - E(\overline{X}_2) - 5 = 0$，备择假设 $H_1: E(\overline{X}_1) - E(\overline{X}_2) - 5 \neq 0$。

(2) 计算统计量

$$T = \frac{\overline{X}_1 - \overline{X}_2 - 5}{s_{D\overline{X}}} = \frac{16.27 - 15.99 - 5}{0.716} = -6.59。$$

(3)既然 T 的绝对值大于 6，可以以非常高的置信度拒绝原假设。《六齐说》关于东周青铜剑比青铜戈戟的锡平均含量高 5 %的记录不能得到实际测量数据的支持。

下面我们尝试检验"东周青铜剑与青铜戈戟的锡平均含量没有差别"的假设：

(1) $H_0: E(\overline{X}_1) = E(\overline{X}_2)$，　　$H_1: E(\overline{X}_1) \neq E(\overline{X}_2)$

(2) 统计量

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s_{D\overline{X}}} = \frac{16.27 - 15.99}{0.716} = 0.39$$

(3) 选 $\alpha = 0.05$，查自由度为 43 + 15 - 2 = 56 的 T 函数表，或用 EXCEL 软件的 TINV 函数，得 $T_{0.025} = 2.003$。

(4) 因为 $T = 0.39 \leqslant T_{0.025} = 2.003$，接受 $H_0: E(\overline{X}_1) = E(\overline{X}_2)$，即实际的测量数据未显示这两种青铜器物的锡平均含量有显著的差别，因此不支持《六齐说》关于青铜戈戟与青铜剑锡含量配方有 5%差别的记录。

实例三　李晓岑(2000)测量了相当数量云南古代铜鼓的铅同位素比值 $^{207}Pb/^{206}Pb$ 和 $^{208}Pb/^{206}Pb$。已知对于铅含量高于 2% ~ 3% 的青铜器，可以利用这两个比值来探索青铜器中铅的矿源。下表列出云南最早的两种类型，万家坝型和石寨山型铜鼓的铅同位素比值的平均值和标准差。这两类铜鼓是在同一地区发现的。

表 7-4　云南万家坝型和石寨山型铜鼓的铅同位素比值数据表

铜鼓类型	数量	$^{207}Pb/^{206}Pb$ 的平均值	$^{207}Pb/^{206}Pb$ 的标准差	$^{208}Pb/^{206}Pb$ 的平均值	$^{208}Pb/^{206}Pb$ 的标准差
万家坝型	5	$\overline{X}_1 = 0.85205$	$s_1 = 0.01508$	$\overline{X}_1 = 2.1040$	$s_1 = 0.01457$
石寨山型	6	$\overline{X}_2 = 0.84933$	$s_2 = 0.00734$	$\overline{X}_2 = 2.1013$	$s_2 = 0.01168$

李晓岑首先检验并接受了铅同位素比值的测量结果服从正态分布的假设，而且对于每一组同位素比值，两类铜鼓测量的方差未见明显差别，即 $\sigma_1^2 = \sigma_2^2$，因此可以进行 t 检验，检验这两种类型的铜鼓的 2 组铅同位素比值是否一致。利用公式(7-4)和(7-5)，对于 $^{207}Pb/^{206}Pb$，计算得到 $T = 0.3925$；而对于 $^{208}Pb/^{206}Pb$，$T = 0.2904$。查自由度为 9 的 t 函数表，取显著性水平 $\alpha = 0.05$，得到 $T_{0.05}(df = 9) = 2.262$。因此，可以以很高的置信度判断，万家坝型和石寨山型铜鼓的 2 组铅同位素比值之间未见明显差异，这两种类型的铜鼓应该有相同的铅矿源。李晓岑对我国西南地区多种类型铜鼓和西南地区一系列铜、铅矿料的铅同位素比值的测量数据做了 t 检验，得出我国西南地区古代铜鼓主要使用当地矿料铸造的推论。

对于两总体的方差 σ_1^2 与 σ_1^2 不一致的情况，可使用 7.6 讨论的非参数假设检验。

7.3 配对样本总体平均值一致性的检验

两个配对样本是指两个样本的成员之间是两两相互配对的，例如研究两代男子身高间的变化，抽取了 n 对父子，父亲组和儿子组两组样本的容量是相等的，都是 n，而且数据对之间是相关的。因此配对样本又称为相关样本。第六章曾提到，大样本条件下对于两个配对样本，可以用 U 检验方法来检验它们的总体平均值的一致性，并给出了计算两个配对样本平均值差的标准差的公式

$$s_{D_{\overline{X}}} = s_{(\overline{X}_1 - \overline{X}_2)} = \sqrt{(s_1^2 + s_2^2 - 2r_{12}s_1s_2)/n} \tag{6-3}$$

式中的 r_{12} 是 X_1 与 X_2 之间的皮尔逊相关系数。为了避免计算相关系数，这里介绍另外一种检验配对样本平均值一致性的方法。两种方法是等效的。

因为两个样本中的元素或实体都是成对的，可以计算每一对观测数值之差

$$D_i = X_{1i} - X_{2i} \tag{7-6}$$

式中 X 的第二个下标 i 是样本中配对实体的编号，$i = 1,2,\ldots n$，而 n 是样本的容量。下一步计算诸 D_i 的平均值 $\overline{D}$ 和标准差 s_D。而平均值 $\overline{D}$ 的标准差为 $s_{\overline{D}} = s_D/\sqrt{n}$。如果 X_{1i} 和 X_{2i} 服从正态分布(这里不一定要求 X_{1i} 和 X_{2i} 的方差相等)，那么统计量

$$T = \frac{\overline{D}}{s_D/\sqrt{n}} \tag{7-7}$$

服从自由度为 $(n - 1)$ 的 t 分布。

下面以作者请两个实验室用中子活化分析方法测量同一批原始瓷片中钾含量的部分数据为例，说明成对样本总体平均值差异的假设检验。表 7-5 列出了测量数据，表的第 2、3 列是 18 对测量数据，组成一组配对样本，第 4 列是成对数据的差值。该表的最右面一列显示差值的正负号，这列数据将在 7.7 节的非参数符号检验中被使用。

表 7-5　两个实验室测量 18 片原始瓷片钾含量的数据

样品编号	钾含量 % 实验室 1 - X_1	钾含量 % 实验室 2 - X_2	差值 % $X_1 - X_2$	差值的符号
1	2.43	2.92	-0.49	-
5	3.37	4.05	-0.68	-
9	3.25	3.48	-0.23	-
13	2.56	2.66	-0.10	-
17	1.86	1.80	0.06	+
21	1.94	1.83	0.11	+
25	1.96	2.08	-0.12	-
29	3.32	3.53	-0.21	-
33	3.2	3.68	-0.48	-
37	3.55	3.51	0.04	+
41	2.13	2.58	-0.45	-
45	2.09	1.87	0.22	+
49	2.09	2.26	-0.17	-
53	1.28	1.43	-0.15	-
57	1.72	1.71	0.01	+
61	3.14	2.75	0.39	+
65	2.67	2.14	0.53	+
69	3	2.51	0.49	+
平均值	2.531	2.599	-0.068	
标准差	0.679	0.782	0.340	

下面检验两个实验室测量钾含量的数据间是否存在系统误差。检验过程如下：

(1) 提出原假设，认为两个实验室测量钾含量的数据一致，实验室间不存在系统误差。

$H_0: E(\bar{D}) = 0$，备择假设为 $H_1: E(\bar{D}) \neq 0$。

(2) 计算统计量

$$T = \frac{|\bar{D}|}{s_D/\sqrt{n}} = \frac{0.068}{0.340/\sqrt{18}} = \frac{0.068}{0.080} = 0.849$$

(3) 选取 $a = 0.05$，查 $df = 17$ 的 t 分布函数，得 $T_{0.025} = 2.110$。

(4) $T = 0.849 < T_{0.025} = 2.110$，接受 $H_0: E(\bar{D}) = 0$，即在 $a = 0.05$ 的显著性水平上没有观察到两个实验室测量陶瓷样品的钾含量的数据间存在系统差异。

另一方面也可以计算 18 对数据的皮尔逊相关系数，得到 $r = 0.901$(计算方法见第九章)。利用公式(6-3) 计算得到

$$s_{D_{\bar{X}}} = s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_1^2 + s_2^2 - 2r_{12}s_1s_2}{n}}$$

$$= \sqrt{\frac{0.679^2 + 0.782^2 - 2 \times 0.901 \times 0.679 \times 0.782}{18}} = 0.080$$

$$T = \frac{|\bar{X}_1 - \bar{X}_2|}{s_{D_{\bar{X}}}} = \frac{|2.531 - 2.599|}{0.080} = 0.849$$

用两种方法计算的 T 值相等，可见这两种检验方法是等效的。

如果将相关样本作为两个独立样本来处理，将会得到不同的检验结果。为了显示这个差别，下面用检验独立样本两总体平均值一致性的方法对这个实例作检验。对于两个独立样本，其计算过程和计算结果为

$$s_{D_{\bar{X}}} = s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_1^2 + s_2^2}{n}} = \sqrt{\frac{0.679^2 + 0.782^2}{18}} = 0.244$$

$$T = \frac{|\bar{X}_1 - \bar{X}_2|}{s_{D_{\bar{X}}}} = \frac{|2.531 - 2.599|}{0.244} = 0.280$$

可以看到，成对样本条件下计算的 $s_{D_{\bar{X}}} = 0.080$，明显小于独立样本条件下计算的 $s_{D_{\bar{X}}} = 0.244$，前者的 T 值也就显著大于后者。因此相对而言，对于同一批数据，作为相关样本处理比独立样本更能检验出总体平均值间微小的差异。

7.4 多个独立样本间总体平均值一致性的检验——一元方差分析(ANOVA)

一元方差分析在英语中称为 One-way ANOVA（Analysis of Variance）。在一些统计软件中用 ANOVA 这个简称。它应用于三个或三个以上样本的总体平均值一致性检验。在考古研究中有时也会提出这类要求，例如在贫瘠、中等和肥沃三种土地资源类型的地区，各调查测量了若干聚落的面积。希望比较三地区之间的聚落平均面积间有没有差异，以了解区域土地资源情况是否影响聚落面积的大小。袁靖(Yuan,2002)曾观测比较山东乳山市翁家埠贝丘遗址三个层位中贝壳的平均宽度，试图探讨人类的超量食用对当地贝类生物期望寿命的影响，这类考古研究实例都可以借助 ANOVA 方法。

7.4.1 一元方差分析的原理和步骤

假设有 k 个样本，每个样本有 n_j 个实体($j = 1,2,\ldots,k$)。总的实体数目 $n_{tot} = \sum_{j=1}^{k} n_j$。第 j 组中第 i 个实体的取值是 X_{ij}，第一个下标表示实体编号，第二个下标表示样本编号。用 $\bar{X}_j$ 表示第 j 个样本的平均值。用 $\bar{X}_{tot} = \frac{\sum_j \sum_i X_{ij}}{n_{tot}}$ 表示总平均值，即各组样本全部实体加在一起的平均值。

下面先定义几个随机量，它们是不同的离差平方和。

(1) 总离差平方和

$$TSS = SS_{tot} = \sum_i \sum_j (X_{ij} - \bar{X}_{tot})^2 = \sum_i \sum_j X_{ij}^2 - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{n_{tot}} \tag{7-8}$$

总离差平方和反映各样本全部个体相对于总平均值 $\bar{X}_{tot}$ 的离散程度。

（2）总组内离差平方和

$$RSS = SS_{wg} = \sum_j \sum_i (X_{ij} - \bar{X}_j)^2 = \sum_j \left(\sum_i X_{ij}^2 - \frac{\left(\sum_i X_{ij}\right)^2}{n_j} \right) \tag{7-9}$$

总组内离差平方和是先计算每个样本的成员相对于本样本的中心 $\bar{X}_j$ 的离差平方和，然后再把各样本的离差平方和加在一起。总组内离差平方和反映各样本内部的离散程度。

（3）组间离差平方和

$$BSS = SS_{bg} = \sum_j n_j (\bar{X}_j - \bar{X}_{tot})^2 = \sum_j \left[\frac{\left(\sum_i X_{ij}\right)^2}{n_j} \right] - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{n_{tot}} \tag{7-10}$$

组间离差平方和是把各样本的元素都“移”到本样本的平均值处，即样本的重心处，然后再计算它们相对于总平均值 $\bar{X}_{tot}$ 的离散程度。组间离差平方和反映样本之间的离散程度。

可以证明总离差平方和等于总组内离差平方和与组间离差平方和之和，即有如下的关系式

$$SS_{tot} = SS_{wg} + SS_{bg} \tag{7-11}$$

这些离差平方和也是统计量，它们的自由度分别是：

总离差平方和的自由度 $df_{tot} = n_{tot} - 1$

总组内离差平方和的自由度 $df_{wg} = \sum_j (n_j - 1) = n_{tot} - k$

组间离差平方和的自由度 $df_{bg} = k - 1$

不难看出总离差平方和的自由度等于总组内离差平方和的自由度和组间离差平方和的自由度之和。即有如下的关系式

$$df_{tot} = df_{wg} + df_{bg} \tag{7-12}$$

还需要定义平均离差平方和的概念，它等于离差平方和被自由度去除。相应有：

平均总离差平方和　$MS_{tot} = SS_{tot}/df_{tot}$　(7-13)

平均总组内离差平方和　$MS_{wg} = SS_{wg}/df_{wg}$　(7-14)

平均组间离差平方和　$MS_{bg} = SS_{bg}/df_{bg}$　(7-15)

一元方差分析应用于检验多个总体平均值的一致性，其基本思路是以平均总组内离差平方和 MS_{wg} 作为标尺去衡量平均组间离差平方和 MS_{bg} 的大小。即以平均的组内离散程度去衡量各样本组平均值 $\bar{X}_j$ 之间的离散程度。为此计算统计量

$$F = MS_{bg}/MS_{wg} \tag{7-16}$$

很显然，在样本数目和每个样本的容量一定的情况下，F 值越大，反映样本平均值之间的差别也越大。可以证明，这个统计量 F 服从自由度为 df_{bg} 和 df_{wg} 的 F 分布。F 分布是根据著名统计学家费舍(Fisher)的姓命名的，它是一个有两个自由度的分布函数，分别称为第一和第二自由度。图 7-1 是 F 分布函数的示意图。

F 函数的取值总是正值，曲线底下的面积等于 1。统计学书中都附有专门的 F 函数

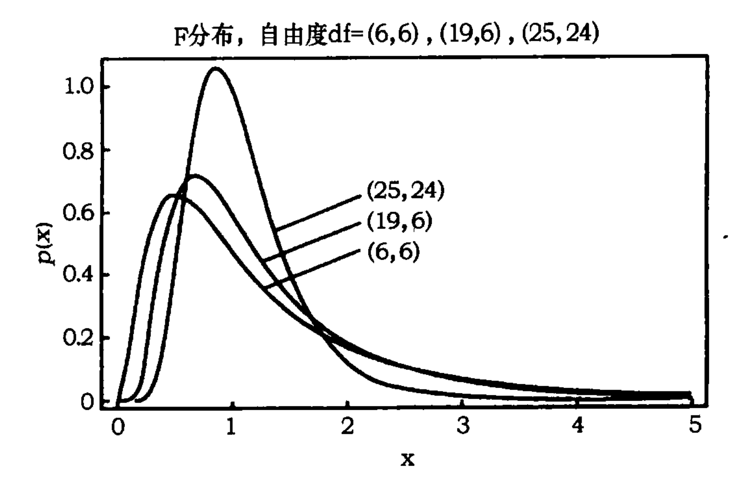


图 7-1 F 分布函数图(引自谢衷洁,2004)

表,给出不同自由度时的 F_α 值,使得 F 大于 F_α 的概率为 α,即 $P\{F > F_\alpha\} = \alpha$。也可以用 EXCEL 软件中的 FDIST($F_\alpha$, $df1$, $df2$)函数,计算 $F > F_\alpha$ 的概率,即 $P\{F > F_\alpha\} = \alpha$。而函数 FINV(概率值 α, $df1$, $df2$)返回 F_α,使得 $P\{F > F_\alpha\} = \alpha$。

7.4.2 ANOVA 实例之一:不同土壤肥瘠程度的地域中聚落平均面积的一致性检验

了解了一元方差分析的基本思想后,回到本节开头提出的,检验三类其土壤肥瘠程度不同的地区之间聚落平均面积有没有差异的实际问题。土壤肥瘠程度分为贫瘠、中等和肥沃三类,表 7-6 的第 2 至 6 列给出分属三类地区 13 个聚落的面积,是进行一元方差分析的原始数据。下表的右面 4 列和最下面的一行给出一元方差分析中间过程的数据。

表 7-6 不同土壤肥瘠程度地域中聚落面积统计和 ANOVA 中间过程数据

土壤肥瘠程度	聚	落	面	积		n_j	$\sum_i X_{ij}$	$\bar{X}_j$	$\sum_i X_{ij}^2$
贫瘠	4	8	7	9		4	24	6	154
中等	17	10	9	12		4	48	12	614
肥沃	20	22	19	9	14	5	84	16.8	1522
总和						13	156		2290

开始检验前,先计算 3 个离差平方和的数值:

$$SS_{tot} = \sum_i \sum_j X_{ij}^2 - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{n_{tot}} = 2290 - \frac{156^2}{13} = 418$$

$$SS_{wg} = \sum_j \left(\sum_i X_{ij}^2 - \frac{\left(\sum_i X_{ij}\right)^2}{n_j} \right) = \left(154 - \frac{24^2}{4}\right) + \left(614 - \frac{418^2}{4}\right) + \left(1522 - \frac{84^2}{5}\right) = 158.8$$

$$SS_{bg} = \sum_j \left(\frac{\left(\sum_i X_{ij}\right)^2}{n_j} \right) - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{n_{tot}} = \frac{24^2}{4} + \frac{48^2}{4} + \frac{84^2}{4} + \frac{156^2}{13} = 259.2$$

验证 $SS_{wg}+SS_{bg}=158.8+259.2=418=SS_{tot}$。说明计算过程中没有错误。这3个离差平方和的自由度分别为12,10和2。

下面进入检验过程。

(1) 假设聚落的平均面积与聚落所在地的土壤肥瘠程度间没有关系。

$H_0:\mu_1=\mu_2=\mu_3$　　　H_1: 至少有一对平均值不相等

上式的 μ_i 表示各类地域聚落的平均面积。

(2) 利用公式(7-16)计算统计量

$$F=\frac{SS_{bg}/df_{bg}}{SS_{wg}/df_{wg}}=\frac{259.2/2}{158.8/10}=8.16$$

其自由度应为2和10。

(3) 用 EXCEL 计算函数 FDIST (8.16, 2, 10) = 0.008,表明 $F\geqslant 8.16$ 的概率为0.008。

(4) 可以在 $\alpha=0.01$ 的显著性水平上拒绝 $H_0:\mu_1=\mu_2=\mu_3$,即认为三种不同土壤类型地区的聚落平均面积是有较明显的差异的。

为了把检验的结果表述的更清楚,一般将上面的方差分析过程和结果总结列于下表。

表 7-7　ANOVA 检验结果汇总表

	df	SS	MS	F	α
组间	2	259.2	129.6	8.16	<0.01
组内	10	158.8	15.88		
全体	12	418			

F(8.16,2,10) = α = 0.008

7.4.3 ANOVA 实例之二:不同葬式墓坑的平均宽度是否有差异

前苏联考古学家克拉斯诺夫调查测量了高尔基州的属公元5—8世纪的“缺水”墓地88座墓葬的葬式和墓坑宽度,表7-8列出原始观测数据(数据引自 Федов-Давыдов (1987))。现用一元方差分析方法判断,不同葬式的墓坑宽度之间是否有差别,或者说墓葬的规模(墓坑宽度)和葬式之间是否有关联。

表 7-8　“缺水”墓地墓葬的葬式和墓坑宽度统计表

墓坑宽度 cm	中心宽度 cm	各葬式墓葬数			Σ和
		普通葬	焚后葬	二次葬	
26—50	38	1	0	0	1
51—75	63	29	7	5	41
76—100	88	21	4	9	39
101—125	113	5	1	3	9
126—150	138	1	0	0	1
151—175	163	0	0	2	2
Σ和		57	12	19	88
平均宽度		77.5	75.5	93.3	80.6

为了减少计算的工作量,在计算各种葬式的墓坑宽度的平均值等统计量时,将墓葬按墓坑的宽度分成6组,每组全部墓葬的墓坑宽度均以该组墓坑宽度的中值替代。第一步,先计算各葬式墓坑宽度的平均值和总平均值:

普通葬 $\bar{X}_1 = \dfrac{(38 \times 1 + 63 \times 29 + \cdots\cdots)}{57} = 77.5$

焚后葬 $\bar{X}_2 = \dfrac{(38 \times 0 + 63 \times 7 + \cdots\cdots)}{12} = 75.5$

二次葬 $\bar{X}_3 = \dfrac{(38 \times 0 + 63 \times 5 + \cdots\cdots)}{19} = 93.3$

全部墓葬 $\bar{X}_{tot} = \dfrac{(38 \times 1 + 63 \times 41 + \cdots\cdots)}{88} = 80.6$

再计算各组的离差平方和和总的离差平方和:

$SS_1 = (38 - 77.5)^2 \times 1 + (63 - 77.5)^2 \times 29 + \cdots\cdots = 19934.25$

$SS_2 = (38 - 75.5)^2 \times 0 + (63 - 75.5)^2 \times 7 + \cdots\cdots = 3125$

$SS_3 = (38 - 93.3)^2 \times 0 + (63 - 93.3)^2 \times 5 + \cdots\cdots = 16723.71$

$SS_{tot} = (38 - 80.6)^2 \times 1 + (63 - 80.6)^2 \times 41 + \cdots\cdots = 42698$

这样组内总离差平方和和组间离差平方和分别为:

$SS_{wg} = SS_1 + SS_2 + SS_3 = 38782.96$

$SS_{bg} = (77.5 - 80.6)^2 \times 57 + (75.5 - 80.6)^2 \times 12 + (93.3 - 80.6)^2 \times 19 = 3924.40$

验证计算过程 $SS_{wg} + SS_{bg} = 38782.96 + 3924.40 = 42707.36$,与 $SS_{tot} = 42698$ 基本相等。因为在这个例子中用墓坑分段宽度的中值替代平均值作近似计算,两种方法计算的总离差平方和之间有一定的误差是容许的。SS_{wg} 和 SS_{bg} 的自由度分别为 $df_{wg} = 88 - 3 = 85$ 和 $df_{bg} = 3 - 1 = 2$。

检验过程如下。

(1) 提出原假设各葬式之间其平均墓坑宽度没有差别,$H_0: \mu_1 = \mu_2 = \mu_3$;相应的备择假设为:至少在二种葬式之间其平均墓坑宽度存在差别。

(2) 计算统计量 $F = \dfrac{3924.4/2}{38783/85} = 4.3$

(3) 设定显著性水平 α,查自由度为2和85的F函数表,并作比较,$F_{0.05}(2,85) = 3.10 < F = 4.3 < F_{0.01}(2,85) = 4.84$。

如果设定 $\alpha = 0.05$,则拒绝原假设;但如果设定 $\alpha = 0.01$,则接受原假设。因此检验的结论是不同葬式的墓坑平均宽度间有一定的差异,或者说墓坑的平均宽度和葬式间存在一定的关联,但关联强度并不高。

7.4.4 ANOVA 实例之三:两周墓葬中青铜容器随葬组合的研究

吴十洲(2001)曾统计研究了两周386座墓葬中青铜容器的数量和组合关系,这些墓葬属不同时代:不同文化地区,墓主人的身份不同。研究的目的是考察两周时期墓葬随葬青铜容器和礼器,特别是用鼎制度有什么规律,与东周文献的记载是否相符,从周初到战国发生了怎样的变化等。因为目前已累积了相当数量关于两周随葬青铜容器的资料,使得吴十洲有可能使用定量方法对此作研究,他使用了一元方差分析和相关分析两种定

量方法。本节将介绍吴应用一元方差分析的部分研究结果,目的在于显示一元方差分析应用于考古课题的过程与作用。关于两周青铜容器的随葬组合研究中应用相关分析的情况将于第九章简单介绍。

吴十洲将326座墓葬按时代早晚分为10段,西周5段、春秋3段和战国2段,根据每一段墓葬中随葬青铜容器数量的分布和平均值,检验10个时代段单个墓葬中随葬青铜容器数量的总体平均值是否相等。提出原假设:各时代段单座墓葬随葬青铜容器数的总体平均值相等。根据平均组间差和平均组内差的比值,对这个样本(10段326座墓葬的随葬品分布)计算得到的 F 值等于1.97,在 $\alpha = 0.05$ 的置信度水平下查自由度(9,316)的 F 函数表,得 $F_{0.05} = 1.91$。因为 $F = 1.97 > F_{0.05} = 1.91$,在 $\alpha = 0.05$ 的置信度水平下拒绝原假设,认为各时代段每座墓葬随葬青铜容器的平均数是有一定差别的,"总的趋势是越向后期发展,随葬青铜容器的平均值越高"。吴还统计了50多座墓主人自铭身份等级的墓葬中青铜容器的数量,他将墓葬分成6个等级,一元方差分析的结论是"墓主自铭身份与墓葬青铜容器的数量之间没有什么差异",他的推论是随葬青铜容器的数量"并不如东周礼书说的那么严格"。需要指出,目前学术界对这些问题是有不同看法的。

7.4.5　关于一元方差分析的前提和分析结果讨论

至今我们仅介绍了一元方差分析的原理与方法,但一元方差分析的应用也需要满足一定的前提条件。它对样本的容量没有要求,可适用于小样本。但是它要求:(1)样本的成员来自服从正态分布的总体,(2)各样本的方差之间差别不显著和(3)抽样是随机的。在实际工作中有的样本的个体数太少,没法检验它们的分布。可以先计算每个样本各实体的离差 $X_{ij} - \bar{X}_j = \sigma_{ij}$,然后把各个样本全部实体的离差 σ_{ij} 合在一起作直方图,或者用后面7.5节将介绍的正态 P-P 图来检验这些离差值 σ_{ij} 是否服从正态分布。关于样本间方差的一致性检验将在7.6节中讨论。但是作为一个经验法则(The rule of thumb),只要样本间最大和最小方差的差别不超过2-3倍,就可以应用一元方差分析。总的说来,一元方差分析还是比较宽容的,容许实际情况对它所要求的前提有所偏离。

一元方差分析的结果可能是接受、也可能是拒绝原假设。当拒绝原假设时,如同7.4.2节实例的情况,这表明诸样本来自总体的平均值之间是有差异的,或者至少有一对样本来自平均值有差异的总体。但推论应该到此为至,一元方差分析本身并不能告诉我们哪一对,或哪几个总体间的平均值有差异。为了进一步在诸 $E(\bar{X}_j)$ 间比较,有的概率统计学家提出了一些方法,如 Turkey's Honestly Significant Difference 方法等。如果样本的数目不是太多,也可以用前面讨论过的两个总体间平均值一致性的 U 检验或 t 检验方法,把样本两两分对来处理。对7.4.2节的实例一,通过两总体平均值的一致性检验,可以推论贫瘠土壤和肥沃土壤环境之间的聚落平均面积有较显著的差别。

7.5　假设检验中对于总体正态分布和总体方差一致性前提的检验问题*

小样本两总体均值一致性的 t 检验和一元方差分析均涉及两个前提条件,要求有关样本服从正态分布,和总体间的方差无显著差别。下面对这两个前提条件的检查或检验

作简单说明。

7.5.1 怎样检查或检验样本是否来自正态分布总体

小样本包含的个体数太少，难以画直方图观察其经验分布。一般情况下，我们掌握的考古学知识也不能给出样本是否服从正态分布的推论。可以利用 SPSS 软件的 descriptive→Explore→Plot 程序，执行柯尔莫哥洛夫-斯米尔诺夫检验和夏比洛-维尔克检验，检验样本的经验分布是否接近于正态分布。其中夏比洛-维尔克检验更适用于小样本的情况（$n \leqslant 50$）。表 7-9 显示 SPSS 软件相关程序对 7.2.2 节的实例一的检验结果，对于男、女墓葬的随葬品数量这两个样本，显著性水平均明显大于 0.10，因此可以接受它们来自正态分布总体的假设。

表 7-9 墓葬随葬品数量分布的正态分布检验（男、女性墓葬分别检验）

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
男性墓葬	0.186	8	0.200	0.949	8	0.705
女性墓葬	0.127	11	0.200	0.957	11	0.738

也可以利用 Normal P-P 或 Normal Q-Q 图来观察或粗略检查样本是否来自正态分布总体。Normal P-P 图是样本的实测累计频率相对于按照正态分布计算的期望累计概率的散点图。如果散点图中的点基本上聚集于一条 45 度对角线附近，那么可以认为样本的经验分布与正态分布差别不大。图 7-2 和图 7-3 分别是 7.2.2 节实例一中男、女性墓葬的随葬品数量的 Normal P-P 图。这个图可以由 SPSS 软件的 Chart 命令产生。

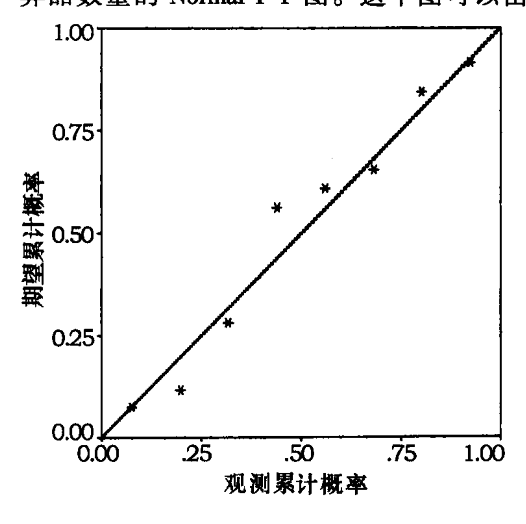


图 7-2 7.2.2 节的实例一中男性墓葬中随葬品数量的 Normal P-P 图

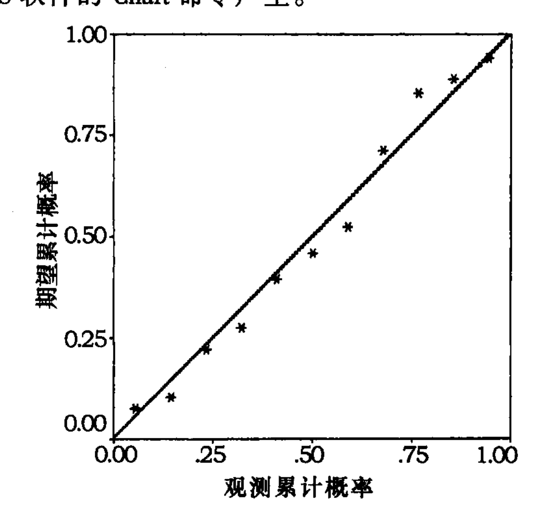


图 7-3 7.2.2 节的实例一中女性墓葬中随葬品数量的 Normal P-P 图

在这两张图中，各点是聚集于 45 度对角线的附近，偏离不大。说明男、女墓葬的随葬品数量分布基本服从正态分布，因此 7.2.2 节实例一进行的对两总体平均值一致性检

验采用 t 检验是合理的。

顺便指出在统计学的应用中，经常要求所研究的样本来自正态总体，包括 7.4 节讨论的一元方差分析等。乃至某些离散型的随机变量的分布也常用正态分布近似。因此，这里介绍的 Normal P-P 对于粗略检验经验分布是否接近正态是有广泛用途的。

7.5.2　两总体方差一致性的检验

两总体小样本的平均值一致性的 t 检验和 ANOVA 的应用都要求样本来自方差相等的总体。本节将讨论两总体间方差的一致性检验。

在 5.3 节中提到，如果总体服从正态分布，则以总体方差 σ^2 为度量尺度的样本离差平方和 $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{s^2(n-1)}{\sigma^2}$ 服从自由度为 $(n-1)$ 的 χ^2 分布。可以证明，对于两个服从正态分布且总体方差一致的样本，其方差的比值 $\frac{s_1^2}{s_2^2}$ 服从自由度为 (n_1-1) 和 (n_2-1) 的 F 分布。因此，对两总体方差的一致性的检验过程如下：

(1) 提出原假设，认为两个总体的方差一致。$H_0:\sigma_1^2 = \sigma_2^2$；备择假设为 $H_1:\sigma_1^2 \neq \sigma_2^2$

(2) 计算统计量 F，并选择 $s_1^2 > s_2^2$

$$F = \frac{s_1^2}{s_2^2} \tag{7-17}$$

(3) 定 $\alpha = 0.05$，查 $F_{0.025}(n_1-1, n_2-1)$

(4) 比较 F 和 $F_{0.025}(n_1-1, n_2-1)$ 的大小，在 $\alpha = 0.05$ 的显著性水平上，作出接受或舍弃 $H_0:\sigma_1^2 = \sigma_2^2$。

需要指出，这实际上是双侧的检验，只是由于技术上的方便，选择 $s_1^2 > s_2^2$，使得 F 总是大于 1。因此选定了 α 后，需要查 $F_{\frac{\alpha}{2}}$ 值。表 7-10 给出对 7.2.2 和 7.3 中三个例子中方差一致性假设的检验过程和结果。

表 7-10　7.2.2 节和 7.3 节总体均值一致性检验三个实例中方差一致性的检验结果

	样本名称	容量 n	标准差 s	F 值	$\alpha = 0.05$ 检出阈
7.2.2 例一	男性墓葬	8	8.17	1.829	3.950
	女性墓葬	11	6.04		
7.2.2 例二	青铜剑	43	2.42	1.117	2.665
	青铜戈戟	15	2.29		
7.3 实例	实验室 2	18	0.782	1.326	2.673
	实验室 1	18	0.679		

三个实例的 F 值均小于相应自由度为 $\alpha = 0.05$ 的检出阈，它们全都通过了两总体的方差一致性检验。说明在这些例子中用 t 检验方法检验样本平均值数学期望的一致性，其方差一致性前提是满足的。

顺便指出，如果使用 SPSS 软件进行两总体均值一致性检验，程序会自动对 $\sigma_1^2 = \sigma_2^2$ 假设作检验，并同时输出 $\sigma_1^2 = \sigma_2^2$ 成立和不成立两种情况下的检验结果。表 7-11 显示 SPSS

软件对 7.2.2 节实例一(男、女墓葬的随葬品数量的均值检验)检验结果的输出(原输出表格稍作删节)。第 3 行和第 4 行分别显示 $\sigma_1^2 = \sigma_2^2$ 成立和不成立两种不同情况下的均值一致性 t 检验结果。

表 7-11 SPSS 软件对 7.2.2 节实例一两总体均值一致性检验的输出表格

F	Sig.		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
0.991	0.333	方差一致	2.804	17	0.012	9.1136	3.25076
		方差不一致	2.669	12.304	0.020	9.1136	3.41501

7.6 两总体平均值一致性的非参数假设检验

在总体方差未知的条件下,小样本两总体平均值一致性 t 检验的前提条件是:两个样本都来自正态总体和两个总体的方差没有显著的差别。但是有时这两个前提条件不成立或难以确认,这时需要用非参数假设检验的方法来检验两样本所属总体的平均值是否有显著差异。非参数检验不涉及总体的平均值和方差等参数,也不需要对总体的分布作什么假设,因此能应用于各层次数据的样本。下面将通过实例来说明非参数检验中的秩和检验和符号检验两种方法。前者适用于独立样本,后者仅适用于不独立的成对样本。

7.6.1 两期聚落面积一致性的秩和检验

表 7-12 列出某地某文化前后两期聚落的面积。早期聚落共 $n_1 = 20$ 个,晚期聚落 $n_2 = 18$ 个。要求 检验两期聚落的总体平均面积是否一致。在表 7-12 中虽然早晚期聚落的面积分别列于上下 2 行中,但前后两期的聚落是按面积的大小统一排序的。这样每个聚落都有一个反映其面积大小次序的序号,或称秩,记录在表 7-12 的第一行中。在对每个聚落面积赋予序号时,需要注意面积相等的聚落。例如早晚期都有一个面积为 56 的聚落,本来它们的序号应该是 9 和 10,现在对这 2 个聚落均赋予序号 9.5。同样原因 3 个面积同为 72 的聚落的序号都被赋予 20,因此不存在序号为 19 或 21 的聚落。

表 7-12 某地某文化前后两期聚落面积的统计表

序号	1	2	3	4	5	6	7	8	9.5	9.5	11	12	13
早期	31	35	40	42	46	50		54	56			61	
晚期							52			56	60		62
序号	14	15	16	17	18	20	20	20	22	23	24	25	26
早期	64	65	67			72		72	73	74			
晚期				68	70		72				75	76	78
序号	27	28	29	30	31	32	33	34	35	36	37	38	
早期				83	84		86				91		
晚期	79	80	81			85		87	88	90		92	

在这个例子中,早期聚落的数目多于晚期的,即 $n_1 > n_2$,我们把实体数少的晚期聚落的序号加起来,$T = 7 + 9.5 + 11 \cdots\cdots + 36 + 38 = 429.5$,$T$ 称为秩和。

可以证明,如果“两期的聚落来自同一个总体”,而且 n_1 和 n_2 都大于 10,那么秩和 T 近似服从平均值和标准差分别为:

$$\mu = \frac{n_2(n_1 + n_2 + 1)}{2} \text{和} \sigma = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}} \tag{7-18}$$

的正态分布。对上面例题计算得到 $\mu = 351, \sigma = \sqrt{1170} = 34.2$。

检验过程如下:

(1) 提出原假设, H_0:两期的聚落来自同一个总体;备择假设 H_1:两期的聚落来自不同的总体。

(2) 计算统计量

$$Z = \frac{T - \mu}{\sigma} = \frac{429.5 - 351}{34.2} = 2.29 \tag{7-19}$$

(3) 选定显著性水平 $\alpha = 0.05$,因为是双测检验,查 $Z_{0.025} = 1.96$。因为 $Z_{0.025} = 1.96 < T = 2.29$,在 $\alpha = 0.05$ 的显著性水平上判断,两期聚落的平均面积是有差别的,从总体上讲聚落的规模随时间有扩大的趋势。

前面提到,要求 n_1 和 n_2 都大于 10,这是秩和检验的前提条件。当这个条件不成立时,秩和是不服从正态分布的。尽管在有的统计学书中,专门列表为小样本的 n_1 和 n_2 值也给出相应的 T_1 和 T_2 值,如果样本的 T 值处于 T_1 和 T_2 之间,则接受原假设,反之,则拒绝原假设。但小样本秩和检验的可信度往往是受到质疑的。

两个独立样本平均值一致性的秩和检验又称为 Mann-Whitney U 或 Wilcoxon rank sum 检验,读者可以使用有关的计算机统计软件来实现。需要指出,非参数假设检验虽然有本节开头所介绍的某些优点,但其检验的功效低,细心的读者会注意到,秩和检验并非是直接检验两样本所属总体的平均值是否一致,而是检验两样本是否来自同一个总体。另外为了保证检验的可靠性,包括秩和检验在内的非参数检验同样要求样本的容量应大些。

7.6.2 两个配对样本平均值一致性的符号检验

前一节讨论的前后两期聚落的面积属于不相关的独立样本,而 7.3 节表 7-5 记录的两个实验室共同测量的 18 片原始瓷的钾含量属于两个配对的样本,因为每片瓷片都有一对钾含量数据。7.3 节使用 t 分布函数检验了它们的平均值的一致性。对于配对样本的总体平均值一致性也可以用非参数方法检验。下面对表 7-5 的例子作非参数的符号检验,同样检验两个实验室的测量数据间是否存在系统误差。由表 7-5 的最右面一列的数据可见, $n = 18$ 个差值中出现正值的次数 $m = 8$,出现负值的次数相应为 $n - m = 10$。

符号检验的过程如下:

(1) 作原假设 H_0:两个实验室的测量数据间不存在系统误差,即两个样本的数据来自同一个总体。如原假设成立,则每片瓷片的一对钾含量数据之差($X_1 - X_2$)的符号可正可负,完全是随机的,而且出现正或负的概率应该是相等的,都等于 0.5。对于 n 片瓷片,差值($X_1 - X_2$)的符号出现 m 个“+”值的概率服从二项式分布(见 4.3.2),而且 $p = q = 0.5$。因此原假设也可写成 $H_0: P\{+\} = P\{-\} = 0.5$。备择假设 H_1 的设定随检验要求是双

侧或单侧而定，本例是检验两个实验室的测量数据间是否存在系统误差，属双侧检验，因此 $H_1: P\{+\} \neq P\{-\} \neq 0.5$。

(2) 下一步在原假设成立的条件下计算 $n = 18$ 个差值中出现正值或负值的次数小于等于 $m = 8$ 的概率。

$$P\{m \leqslant 8 \text{ 或} (n - m) \leqslant 8\} = 2\left(\sum_{m=0}^{8} C_{18}^{m} (0.5)^{18}\right) = 0.815$$

(3) 判断：两个实验室的 18 对数据中，正负符号次数的差别大于等于 2 的概率为 81.5%，属高概率事件，因此可以以很高的置信度接受实验室间不存在系统误差的原假设，即 $H_0: P\{+\} = P\{-\} = 0.5$。

在本实例中样本容量 $n = 18$，属小样本。当样本中配对实体的数目超过 30 时，二项式分布接近正态分布，数据的处理将更方便简易，这将在第八章中详细介绍。

符号检验只考虑差值的符号而没有考虑差值的大小，虽然它简单明了，但并没有充分利用数据中更多的信息。非参数检验中的 Wilcoxon 符号秩和检验既考虑了差值的符号又考虑差值的大小。鉴于本书的篇幅和符号检验在考古研究中应用的有限性，这里不作介绍，有兴趣的读者可参考有关的统计学书籍。

第八章　总体比例数的估计和假设检验

前面几章我们讨论了总体平均值和方差等参数的估计问题和假设检验。但在考古研究中还经常碰到样本和总体比例数的问题,例如:(1)根据墓地男女人骨数的比例判断墓地所属氏族的男女性比是否正常;(2)一位旧石器考古学家在某地区随机采集石制品,其中有燧石制品。他当然会用实际采集的燧石质制品的百分比作为该地区石制品中燧石质制品所占百分比的估计量。他希望估计的置信度达95%,而反映精密度的估计误差不高于10%,那么这位考古学家至少应采集多少件石制品;(3)墓地甲发掘墓葬100座,其中60座带有随葬品,而墓地乙发掘墓葬50座,其中带有随葬品的墓有35座,希望判断这两个墓地所属氏族在墓葬制度带不带随葬品的比例上有无明显差别。上面3个例子都涉及比例数的问题,而这类关于比例数的问题都是与二元变量有关,即涉及某种二元属性的取值,如性别的男女,石制品的石质是否是燧石,墓葬带不带随葬品等。我们在4.3.2中已看到,作为贝努利试验结果的二元变量服从二项式分布,因此要用二项式分布来处理上述的问题。第四章还提到,当样本的容量较大,$n>30$时,二项式分布接近于正态分布,用正态分布来处理可以极大地简化计算过程。下面通过实例来讨论总体比例数的估计和假设检验。

8.1　单总体比例数的假设检验:检验墓地人骨男女性比是否正常

山西夏县东下冯遗址的龙山墓地共发掘出17具成年人骨,其中男性11具,女性6具。可以计算,这批人骨的性比值R(男性人数/女性人数)=11/6=1.83。由于生男生女的概率是基本相等的($p=q=1/2$),正常情况下人群的性比应该接近于1。这个样本观测到的性比值明显偏离正常值“1”,能否由此判断,东下冯遗址龙山时期所有埋葬的成员(总体)的性比也偏离正常值,即实际观测到的性比值偏离正常值“1”是属于随机涨落,抑或当时当地所埋葬的全部成员的性比就是异常的,系男多女少。为此要对总体的性比作假设检验。检验过程如下:

(1) 提出原假设 H_0: 东下冯龙山时期所有被埋葬的成年成员的性比是正常的,即 $p=q=0.5$;备择假设为 $H_1:p\neq q\neq 0.5$。

(2) 已知 n 个个体中出现 k 个是男性的概率服从二项式分布${}_nC_kP^kq^{n-k}$。在原假设成立的前提下,利用 二项式分布计算在一个17具人骨的随机样本中男性占11具以上(包括11具)的概率(为了计算方便,可以计算女性人骨少于6具[含6具]的概率)。

$$P\{n=17,\text{男性}\geqslant 11\}={}_{17}C_0P^{17}+{}_{17}C_1P^{16}q+{}_{17}C_2P^{15}q^2+\cdots\cdots+{}_{17}C_6P^{11}q^6=0.166$$

(3) 无论选 $\alpha=0.1$ 或 0.05, $P\{n=17,\text{男性}\geqslant 11\}$ 均大于 α,因此都应该接受 H_0, 即不能因为实际观测到的样本的性比明显偏离正常值“1”,达1.83,而推断东下冯龙山时期埋葬的成年人骨总体上性比异常。即使在显著性水平 $\alpha=0.16$ 水平上,也不能推断总体

的性比不正常。

研究我国新石器时代墓地人骨性比的另一个例子是陕西华县元君庙仰韶墓地。经性别鉴定的成年人骨 146 具,其中男性 85,女性 61,性比 R = 85/61 = 1.37。同样希望判断元君庙仰韶墓地总体的人骨性比是否正常。元君庙墓地的人骨材料数大于 30,属大样本。因此男性人骨出现次数 k 这个随机变量所服从的二项式分布十分接近 $\mu = np, \sigma^2 = npq$ 的正态分布 $N(\mu, \sigma^2)$,从而可以用正态分布来判断总体性比是否正常。计算和判断过程中使用频率比频次更为方便,为此引入 n 次观测中男性人骨出现的频率$\left(\hat{p} = \frac{k}{n}\right)$这个随机变量,$\hat{p}$ 同样接近正态分布,它的数学期望应该是 p,而标准差是$\sqrt{\frac{pq}{n}}$。这样变量

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}} \tag{8-1}$$

服从标准型的正态分布。

下面利用正态分布对元君庙墓地的性比进行检验。

(1) 提出原假设 H_0: 元君庙墓地人骨的性比正常,即 $p = q = 1/2$,备择假设 H_1:人骨的性比不正常,即 $p \neq q \neq 1/2$。

(2) 计算统计量:

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{85/146 - 0.5}{\sqrt{0.5 \times 0.5/146}} = 1.986。$$

(3) 选取 $\alpha = 0.05$, 查正态分布函数表 $Z_{\frac{\alpha}{2}} = 1.96$。

(4) 因为 $Z > Z_{\frac{\alpha}{2}}$,在 $\alpha = 0.05$ 的水平上,拒绝原假设,接受备择假设,即在 $\alpha = 0.05$ 的显著性水平上,判断元君庙墓地的总体人骨性比偏离正常性比值。

对比东下冯龙山墓地和元君庙仰韶墓地,尽管实际观测的东下冯样本的性比值(1.83)比元君庙样本的性比值(1.37)高出很多,但是假设检验的结论却是相反的。同样在 $\alpha = 0.05$ 的水平上,检验结果认为元君庙墓地的总体人骨性比偏离正常值,却接受了东下冯龙山墓地总体性比正常的假设。看起来这似乎有悖于常识。产生这种情况的原因是因为东下冯龙山墓地的人骨数太少。两个检验的可靠性是不一样的。关于样本的容量在总体比例数的假设检验中对犯两类错误的概率以及对总体比例数的估计中可信度和精密度的影响,将在下一节中详细讨论。不过东下冯和元君庙的例子清楚表明,简单根据样本性比的观测值直接去推断总体的性比是否正常是很危险的,特别是小样本的情况。必须根据性比服从二项式分布的知识进行统计推断。本书作者(1990)曾系统统计了 1989 年以前发表的我国 32 处墓地的人骨性比,并由此对这些墓地人骨的总体性比进行统计推断,发现 16 个墓地性比异常。其中除巫山大溪墓地外,都是男性过半。北首岭、半坡、姜寨和史家村等 4 个陕西的仰韶墓地,样本的性比值出现的概率小于 0.01,人骨性比属高度异常。例如半坡墓地 62 具人骨中,男性有 52 具。陕西仰韶墓地人骨性比高度异常的原因,是当时的埋葬制度所导致,还是由于其他原因导致仰韶氏族成年人口本身的性比就不正常,男性多于女性,这值得进一步研究。

8.2　单总体的比例数的估计中置信度、精密度和样本容量三者间的关系

本节根据一个实例来讨论估计总体比例数时置信度、精密度和样本容量三者间的关系。有一位旧石器考古学家在某地区随机采集了200件石制品，其中有燧石制品50件。根据这组观测数据，他能以多高的置信度和精密度对该地区燧石制品的比例数作估计。

对于这个样本，燧石制品出现的频率是 $\hat{p}=\frac{50}{200}=0.25$。他当然会用这个频率值 $\hat{p}=0.25$ 作为该地区石制品中燧石质制品所占的百分比 p 的点估计量，用 $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}=\sqrt{\frac{0.25\times0.75}{200}}=0.031$ 估计 p 的标准差。如果要求对 p 的区间估计的置信度为 $(1-\alpha)$，那么估计区间应该是：$\left[\hat{p}-Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$。可以计算这个区间估计的精密度。估计的精密度与估计的相对误差成反比，后者定义为估计区间的半宽度被区间的中心值去除所得的商值，即

$$\text{总体比例数估计的相对误差}=\frac{Z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}}{\hat{p}} \tag{8-2}$$

如果这位考古学家希望估计的置信度为95%，这样对该地区燧石质制品的比例数 p 的估计区间是 $[0.250\pm1.96\times0.031]$，即 $[0.250\pm0.061]$。估计的相对误差为 $\frac{0.061}{0.25}=0.24$。可见估计的精密度并不高。

公式(8-2)显示，对总体比例数区间估计的置信度和精密度是相互制约的，置信度越高，α 值越小，$Z_{\frac{\alpha}{2}}$ 越大，则估计区间越宽，估计的相对误差也越大，从而估计精密度越低。这种关系也反映在下表所列出的这位旧石器考古学家实际样本的数据中。表8-1显示了这个实例中对应不同的 α 值时的上、下置信阈，置信区间宽度，估计的置信度和相对误差。表中从上到下，估计的置信度不断增加，但置信区间宽度增加，估计的精密度却不断降低。

表8-1　对应于不同的 α 值时估计总体比例数的置信区间宽度、置信度和精密度间的关系

α	$Z_{\frac{\alpha}{2}}$	置信区间上阈	置信区间下阈	置信区间宽度	置信度 $(1-\alpha)\%$	精密度指标相对误差
0.2	1.28	0.210	0.290	0.080	80	0.169
0.1	1.64	0.199	0.301	0.102	90	0.204
0.05	1.96	0.189	0.311	0.122	95	0.243
0.01	2.58	0.170	0.330	0.160	99	0.320

由公式(8-2)可见，只有增加观测量 n，才能同时提高估计的置信度和精密度，当然这是以多支出研究经费、精力和时间为代价的(见5.3)。如果这位考古学家在对该地区燧石制品总体比例数的估计中仍希望保留95%的置信度，但要求估计的相对误差不大于

5%,那么他至少应采集多少件石制品(n)呢。利用公式(8-2)计算

$$\frac{Z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}}{\hat{p}} = 0.05$$

$$1.96\sqrt{\frac{0.25(1-0.25)}{n}} = 0.25 \times 0.05$$

$$n = 4610$$

答案是至少需要随机采集 4610 件石制品,才能使得对该地区燧石制品比例数的区间估计同时达到 95%置信度和 5%相对误差的要求。

8.3 两个总体比例数一致性的假设检验

前两节讨论了单总体比例数的估计和假设检验,本节将通过实例来讨论两个总体比例数一致性的假设检验。已知墓地甲发掘了 100 座墓葬,其中 60 座带有随葬品,而墓地乙发掘了 50 座墓葬,其中 35 座带有随葬品。希望根据这两个样本来判断,两墓地带随葬品墓葬的比例数是否一致,即这两个墓地所属氏族在是否带随葬品方面有没有差异。这个实例是要利用两个样本中带随葬品墓葬比例数的差($\hat{p}_1 - \hat{p}_2$),去推断总体相应比例数的差($p_1 - p_2$)是否等于零。在进行检验前,需要先计算随机变量($\hat{p}_1 - \hat{p}_2$)的标准差 s_{Dp},计算公式如下:

$$s_{Dp} = \sqrt{\frac{\overline{p}\,\overline{q} \times (n_1 + n_2)}{n_1 n_2}} \tag{8-3}$$

式中 $\overline{p}$ 和 $\overline{q}$ 分别是两个样本的 $\hat{p}_i$ 和 $\hat{q}_i$ 值的计权平均,n_1 和 n_2 是两个样本的容量。

已知 $\hat{p}_1 = 0.6, \hat{q}_1 = 0.4, \hat{p}_2 = 0.7, \hat{q}_2 = 0.3$,计算

$$\overline{p} = \frac{0.6 \times 100 + 0.7 \times 50}{100 + 50} = 0.633$$

$$\overline{q} = 1 - \overline{p} = 0.367$$

所以,$s_{Dp} = \sqrt{0.633 \times 0.367 \times \frac{100 + 50}{100 \times 50}} = 0.0835$

检验过程如下:

(1) 提出原假设 H_0:两墓地带随葬品墓葬的比例数无差别,即 $p_1 = p_2$;$H_1: p_1 \neq p_2$。

(2) 计算随机变量

$$Z = \frac{|(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)|}{s_{Dp}} = \frac{|(0.6 - 0.7) - 0|}{0.0835} = 1.20$$

(3) $Z = 1.2$ 查双侧情况下的正态函数表,得 $\alpha = 0.230$。

(4) 无论选择显著性水平 $\alpha = 0.05$ 或 $\alpha = 0.1$,均小于 0.23,因此都应该接受原假设 H_0,即不能认为两墓地之间带随葬品墓葬的比例数有明显差别。

如果检验结果拒绝原假设 H_0,则可以进一步对两个总体的比例数差($p_1 - p_2$)作区

间估计并讨论估计区间的置信度和精密度等。

8.4　用"子弹形"图比较多个总体比例数的差异：以分析赤峰考古调查资料为例

子弹形图是根据英文"Bullet graph"翻译的，它能较为直观地显示两个或多个总体间平均值和比例数的差异，以及数据的离散程度。这里我们将通过赤峰中美联合区域考古调查阶段性报告的资料来介绍应用子弹形图显示多个总体比例数差异的比较。

赤峰中美联合考古研究项目(2003)于1999和2000年在赤峰地区区域考古调查中记录统计了1691个有地面考古遗存的采集点，每个采集点的面积是1公倾。在这些采集点上共采集了24510件陶片，其时代跨度从兴隆洼期一直到辽代。在每个采集点上采集到的陶片数目当然是不等的，在其中的282个采集点，陶片的分布较为稀疏，在1公倾的面积上采集到的陶片数少于5片，研究者称之为"小采集"点。出自小采集点的陶片数共有476片，占总陶片数的1.9%。研究者对小采集点的出现原因作了讨论，认为"如果这些稀疏散布的陶片是由古代遗物的近期搬运所致，我们应当预期这种搬运对每个时期的陶片都有类似的影响"。但子弹形图明显地显示各时期"小采集"陶片的比例数偏离平均值1.9%的情况是很不一样的，从而排除了"近期搬运所致"的可能。表8-2列出赤峰地域调查中所采集的各时期的总陶片数和各时期小采集的陶片数。

表8-2　赤峰地区采集的各时期的陶片数和小采集的陶片数

时代和文化类型	所有采集 陶片数量	少于5片的小采集 陶片数量	少于5片的小采集 百分比%
兴隆洼	55	3	5.5
赵宝沟	263	6	2.3
红山	1546	18	1.2
小河沿	178	4	2.2
夏家店下层	7495	117	1.6
夏家店上层	6732	115	1.7
战国-汉	2983	52	1.7
辽代	5258	161	3.1
总计	24510	476	1.9

根据表8-2的数据和本章前面关于样本比例数所服从的分布的知识，可以计算各时期"小采集"陶片总体比例数不同置信度的估计区间。图8-1是一张子弹形图，它显示了从兴隆洼文化到辽代各时期"小采集"陶片的总体比例数的置信区间，其置信度分别为80%、95%和99%。

从图上看到对于兴隆洼、赵宝沟和小河沿文化，因为样本的陶片数太少，其总体"小采集"陶片比例数的估计区间极宽，与各时期小采集陶片的平均比例数(1.9%)的比较没有统计学的意义。例如，尽管兴隆洼小采集陶片的比例数最高，达5.5%，偏离平均值1.9%甚远。我们应主要考虑红山、夏家店下层和辽代3个时期的小采集陶片数，因为其

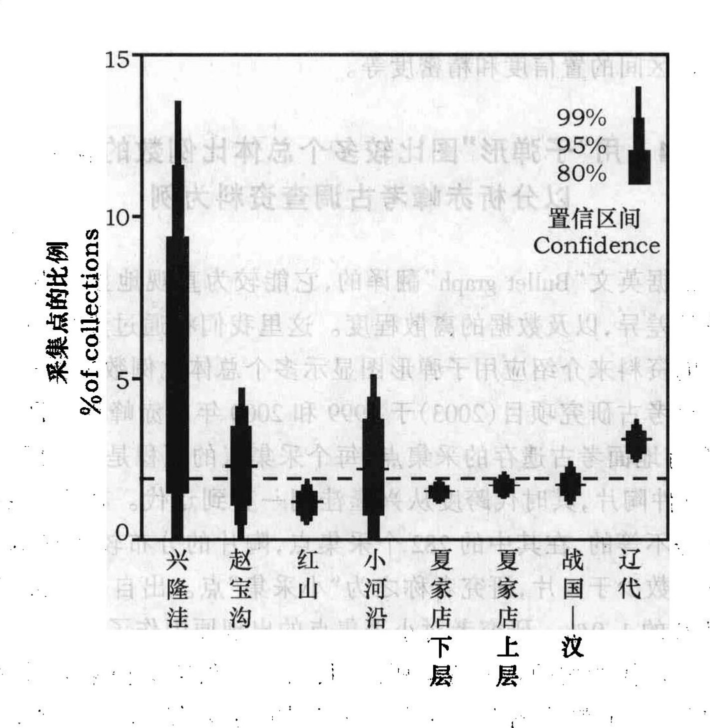


图 8-1 赤峰地区各时期"小采集"陶片的总体比例数的子弹形图，标出了置信度分别为 80%、95%和 99%的估计区间，图上虚线表示全部"小采集"陶片的总体比例数 1.9%

数量较多。这 3 个时期的"小采集"比例数 99%置信度的估计区间均不与代表各时期平均比例数 1.9%的虚线重叠。因此有很大的把握认为辽代小采集陶片的比例数偏高，以及红山文化和夏家店下层文化小采集陶片的比例数偏低，均明显偏离于平均比例数 1.9%。赤峰考古调查的研究者们认为，各时期小采集陶片的比例数的差异可作为这些稀疏散布的各时期陶片不是近期农业活动所致的一个证据。他们认为这些小采集陶片是有考古意义的，为此他们还对小采集陶片的资料作相关分析来加强这一论据，我们将在第九章讨论相关和回归时再回到赤峰小采集的例子。

总之，从上面的实例看到子弹形图在比较几个总体的比例数时，显示出高度的直观性和形象性。子弹形图在表述考古资料中，包括对总体间平均值的比较中也得到广泛的应用。

8.5 考古调查中某类实体的缺失是否说明该类实体确实不存在

假设某地区的考古普查中未见到某种动物，或未见到某种材质的石器，能否就此推断该地区当时的确不存在该种动物或该类石质的制品呢。这里需要非常小心，因为存在另一种可能性。如果该种动物或该类石质制品在总体中的比例数 p 很低，而采集的样本容量 n 又不大，那么有相当大的可能，在采集的样本中不出现该种动物或该类石质制品。正好似随机抽取几张扑克牌，完全有可能其中未出现"A"，但不能由此推断，整付扑克牌

中不包含“A”。如果在某地区随机采集了 n 件石制品未见燧石制品，虽然我们不能肯定该地区不存在燧石制品，但是我们可以问该地区燧石制品的总体百分比小于1%或小于5%的概率多大。这个问题也可以反过来提问，即如果燧石制品的百分比小于1%或小于5%，随机采集的 n 件石制品未见到燧石制品的概率多大。两种提问的方式虽不一样，但实质是一致的。而对于后面一种提问方式，概率值的计算较为方便，它可以用二项式分布的第一项 $C_n^0 p^0 (1-p)^n$ 来计算。如果 $n = 1$ 和 $p = 0.01$，那么未见燧石制品的概率是0.99。这时只有 $(1-0.99) = 0.01$ 的置信度判断总体燧石制品的比例不大于1%。如果 $n = 10$ 和 $p = 0.01$，那么未见燧石制品的概率是 $(0.99)^{10} = 0.904$，将有 $(1-0.904) = 0.096$ 的置信度判断总体燧石制品的比例数不大于1%。当 $n = 100$ 时，将有0.634的置信度判断燧石制品的总体比例数不大于1%。该类问题的一般表达式是

$$\text{判断燧石制品的总体比例数不大于 } p \text{ 的置信度} = [1-(1-p)^n] \tag{8-4}$$

如果要求判断 $p \leqslant 0.01$ 的置信度达95%，则有

$$0.95 = (1-0.99^n)$$

解这个方程，得 $\log 0.05 = n\log 0.99$，$n = 298$，即如果在298件石制品中未发现燧石制品，那么可以有95%的置信度判断总体燧石制品的比例不大于1%。表8-3列出，不同的样本容量 n 时，判断感兴趣事件在总体中不同的出现概率 p 的置信度。例如当 $n = 150$ 时，未见某种实体，则该种实体的总体比例数低于2%的概率为95%，低于1%的概率为78%，低于0.1%的概率为14%。

表8-3　不同容量的样本中未见某种实体时，判断该种实体在总体中的比例数低于某值 p 的置信度计算表

样本容量 n	$p = 0.1\%$	$p = 0.5\%$	$p = 1\%$	$p = 2\%$	$p = 5\%$
20	0.02	0.095	0.182	0.332	0.642
30	0.03	0.14	0.26	0.455	0.785
50	0.049	0.222	0.395	0.636	0.923
100	0.095	0.394	0.634	0.867	0.994
150	0.139	0.529	0.779	0.952	>0.999
200	0.181	0.633	0.866	0.982	1
400	0.33	0.863	0.982	>0.999	1
700	0.504	0.97	0.999	1	1
1000	0.632	0.993	1	1	1
1600	0.794	>0.999	1	1	1
3000	0.95	1	1	1	1

由表8-2可见，在实际样本中未见某类实体的情况下，样本的容量越大，估计该类实体的比例数低于一定数值的置信度愈高，而当样本容量一定时，要求估计的比例数越低，估计的置信度也越低。

第九章　两个数值变量之间的关系——相关与回归

前面几章讨论了考古实体按单个随机变量的描述性统计，讨论了单个数值型随机变量的参数估计和假设检验。但是很多情况下需要同时考虑考古实体的两个或两个以上的数值型属性，这 时除了要分析实体按两个数值变量的分布、每个变量的平均值和方差外，还要考虑两个变量相互之间的关系，本章将介绍数值变量间的相关分析和回归分析。

9.1　实体按两个数值变量经验分布的图形表述——散点图

第三章曾介绍用直方图、茎叶图和箱点图等来形象地描述实体按单个数值属性的分布，在这些图上可以直观地看到实体按某属性分布的中心位置、分布宽度以及是单峰还是双峰分布等。为了直观地观察实体同时按两个数值属性的分布，需要用散点图。散点图分别以实体的两个属性作为 X 和 Y 坐标轴，根据每个实体两个属性的取值决定它在 XY 为坐标轴的平面上的位置。在散点图上不仅能看到实体分别按这两个属性的分布特征，包括实体分布的中心、分布宽度、实体的分组，此外散点图还可以显示两个属性之间的关系。图 9-1a 到图 9-1d 是 4 个散点图的例子。

图 9-1a 显示辽宁西部发现的 37 把东周时期青铜剑按锡和铅的百分含量两个属性分布的散点图。从图上观察，剑的锡和铅含量似乎是随机的，在 0～28%间波动。无论是对于剑的分布，或锡和铅之间的关系，似乎都看不到什么明确的规律。

图 9-1b 显示 18 片原始瓷的钾含量在两个中子活化分析实验室测量结果的散点图，这是 7.3 节中的例子的数据。该图中代表 18 个瓷片的点基本上形成一条接近座标原点

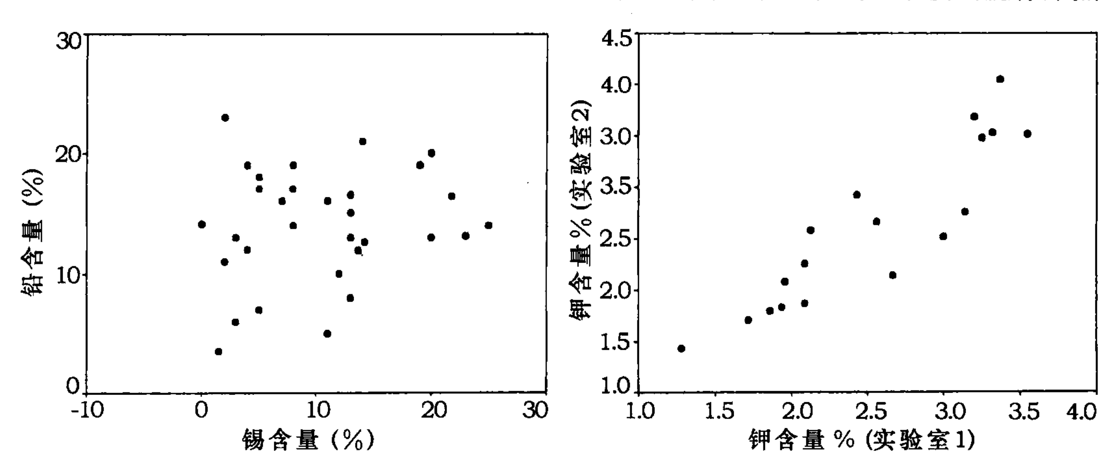


图 9-1a　辽宁西部发现的 37 把东周青铜剑按锡和铅百分含量分布的散点图(部分实验点重叠)

图 9-1b　两个中子活化分析实验室分别测量 18 片原始瓷片中钾含量结果的散点图

的直线，说明同一个瓷片两个变量的取值 X 和 Y 同时高或同时低，即反映两个实验室测量数据的基本一致性。这与第七章"在 $\alpha = 0.05$ 的显著性水平上没有观察到2个实验室测量陶瓷样品的钾含量存在明显的系统差异"的假设检验的结论相符。实验点基本组成直线的情况表明 X 和 Y 的线性相关性，也正是本章9.2节所要讨论的问题。

图9-1c是本书作者测定的商周时期多个地点出土的57片原始瓷，按其Cr和Ce两元素含量分布的散点图。该图显示这些瓷片基本上可分成高Cr低Ce、低Cr低Ce和低Cr高Ce三类。进一步的研究可以将瓷片的这种分类与瓷片的出土地点相对应。另外该图还显示这些瓷片的Cr与Ce含量间不存在明确的相关关系。

图9-1d显示一个假想的样本中18个实体按其两个数值变量分布 X 与 Y 的散点图。无论 X 的取值如何，Y 的取值总是在16—18.5间小幅度的波动。也就是说 Y 的取值是独立的，不依赖于 X 的数值。

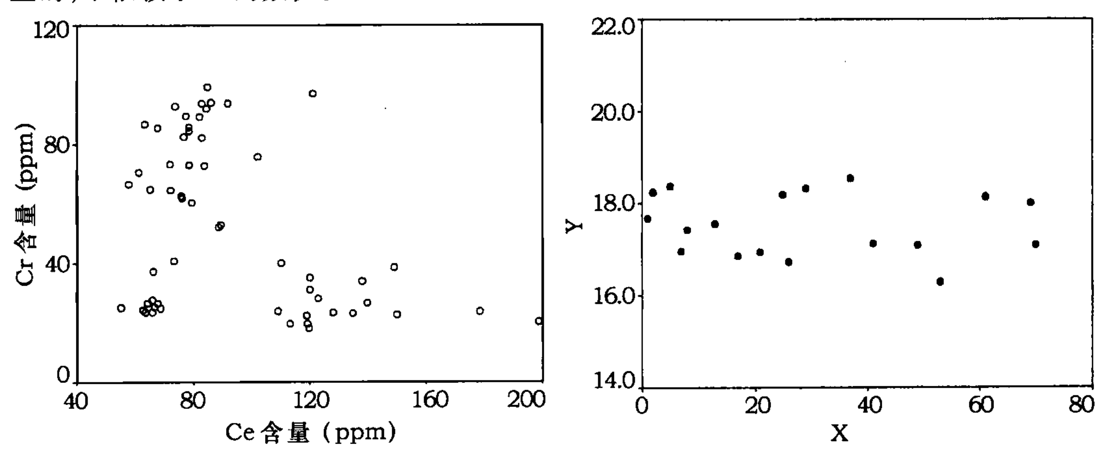


图9-1c　57片原始瓷片相对于它们的Cr和Ce元素含量分布的散点图

图9-1d　假想样本中18个实体按其两个数值变量X与Y分布的散点图

上面4张图表明，散点图能直观地揭示实体按所选两属性的分布规律，并对实体进行分类，同时还能揭示两属性间的关系。散点图在各个学科，包括考古学中得到广泛的应用。本章的内容将局限于两个属性间的相关关系，主要是线性相关关系的讨论。

9.2　线性回归的基本原理和皮尔逊相关系数

相关分析和回归分析是研究两个数值变量之间的关系，但这种关系不是我们所熟悉的函数关系 $Y = f(X)$。在函数关系中，自变量 X 一般能唯一地决定应变量 Y 的取值。回归分析所研究的变量间的关系并不是这种完全确定的关系，而是一种相关关系。这里当 X 确定后，Y 的取值仍可以在一定的范围内波动，Y 的取值分布经常接近于正态分布。例如父母的平均身高并不能绝对确定子女的身高，后者还受遗传过程中的随机因素和后期的营养条件等影响。但在一般情况下，父母的平均身高对子女的身高是有很大的影响的，两者间是相关的。在社会现象，包括考古现象中，完全确定的函数关系是少见的，更多的是各种变量之间的相关关系。例如一个地区新石器时代的聚落大致是年代越晚聚落面积越

大,但具体到某个聚落,它的年代并不能确切地决定该聚落的面积,聚落的年代和面积间存在的是一种相关关系,而不是函数关系。下面将通过考古实例来介绍直线回归的基本原理和皮尔逊相关系数。

设某地曾有一个生产 A 型彩陶的中心,其产品输出到邻近地区。在离该生产地点不同距离 X_i 公里的一些同时代的遗址中发现了 A 型彩陶陶片,并统计了各地每立方米文化堆积中 A 型彩陶片的平均数目 Y_i。统计结果列于表 9-1 和图 9-2。

表 9-1 12 个遗址离 A 型彩陶的生产中心的距离和发现的 A 型彩陶残片的密度

遗址号	1	2	3	4	5	6
离 A 型彩陶生产点距离(公里),X_i	4	7	15	20	21	24
每立方米堆积中 A 型彩陶片平均数,Y_i	95	84	89	67	42	66
遗址号	7	8	9	10	11	12
离 A 型彩陶生产点距离(公里),X_i	28	29	33	35	36	44
每立方米堆积中 A 型彩陶片平均数,Y_i	38	8	40	35	56	38

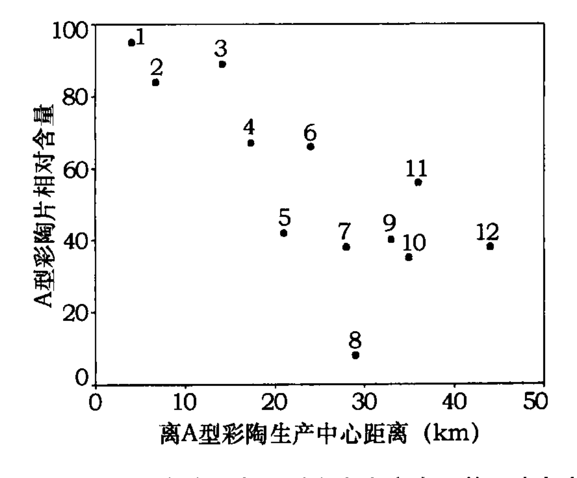


图 9-2 12 个遗址离 A 型彩陶生产中心的距离与其文化堆积中 A 型彩陶片密度间的关系图

从表和图看出,随着遗址离 A 型彩陶生产中心的距离增加,其堆积物中 A 型彩陶片的相对数量总体上呈下降趋势。而且除个别点外,下降趋势接近于线性的。因此很自然地希望用线性下降的规律来描述实际的下降趋势,或者说用一条由方程 9-1 所描述的直线来拟合上述 12 个实验点。

$$\hat{Y} = bX + a \tag{9-1}$$

在未建立这个直线方程时,我们只能用这 12 个遗址每立方米堆积物中 A 型彩陶数目的平均值和标准差 $\overline{Y} = 54.8 \pm 16.0$ 来预测每个遗址的 A 型彩陶密度,预测是很不精确的。如果能够建立这个直线方程,那么知道了某遗址离生产中心的距离 X_i,就可以更准确地预测该遗址每立方米堆积物中 A 型彩陶的数目 $\hat{Y}_i$。问题在于怎样找一条最佳的直线,即怎样来定直线方程(9-1) 的截距 a 和斜率 b。

9.2.1　线性回归方程的参数 a 和 b 的确定

在线性回归中一般把 X_i 当作自变量。任意选定一组 a 和 b 值，即任意选定一条直线后，对于每个 X_i，可以用公式(9-1)计算 $\hat{Y}_i$，称为对应于 X_i 的预测值。实际测量值 Y_i 与预测值 $\hat{Y}_i$ 之间的差$(Y_i - \hat{Y}_i)$称为残差。它是从实际测量点(X_i, Y_i)平行于 Y 轴作直线与拟合直线相交所形成的那段线段的长度(见图 9-3)。残差的大小因 a 和 b 取值的不同而变化，可能为正也可能为负。残差的大小反映了直线拟合程度的优劣，最佳的拟合直线当然应该使得全部实验点都尽量地接近该直线，即全部残差的绝对值都尽量小。为了定量地描述直线拟合程度的优劣，定义残差平方和 RSS：

$$RSS = \sum (Y_i - \hat{Y}_i)^2 \tag{9-2}$$

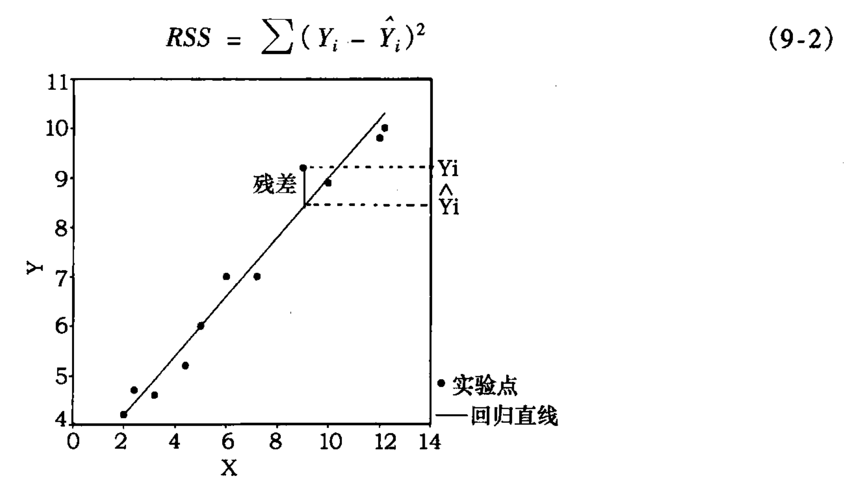


图 9-3　线性回归分析中残差的定义

残差平方和的数值反映了拟合直线接近所有实验点的程度。选择 b 和 a 的标准，即最佳拟合的标准应该是使得残差平方和最小。可以证明，按照下面两个公式计算确定 b 和 a 后，所得的 RSS 最小。

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{L_{xy}}{L_{xx}} \tag{9-3}$$

$$a = \bar{Y} - b\bar{X} \tag{9-4}$$

公式(9-3)中的 L_{xx} 和 L_{xy} 称为内积系数，L_{xx} 和 L_{xy} 除以$(n-1)$就是 X 的方差 σ_X^2 和 X，Y 间的协方差 $\mathrm{cov}(x, y)$，n 为样本的容量。公式(9-4)说明最佳拟合直线是通过所有数据点(X_i, Y_i)的重心$(\bar{X}, \bar{Y})$。这样得到的最佳拟合直线称为回归直线，b 是回归直线的斜率，又称为回归系数，a 是回归直线的截距。回归直线是一条离差平方和最小的拟合直线。

对于上面 A 型彩陶片按遗址距离分布的例子，用公式(9-3)和(9-4)计算得到 $b = -1.64$，和 $a = 95.40$。相应回归直线的方程为

$$\hat{Y} = -1.64X + 95.4 \tag{9-5}$$

b 值为负，说明当 X(距离)增加时，Y 值(每立方米堆积物中 A 型彩陶片的数目)减少。现在有各种计算机软件来计算这些回归方程的参数，不用再花很多的时间来人工计算了。

9.2.2 线性回归方程的检验

上面讨论了怎样计算 a 和 b 以便得到最佳的拟合直线，得到回归方程，但是并没有涉及一个重要的问题，即所得到的回归方程是否有意义，置信度有多高。对于任何一个样本 $(X_i, Y_i)(i = 1,2\cdots n)$ 都可以建立一条回归直线，但只有当样本的诸 (X_i, Y_i) 数据对之间客观存在的关系接近于线性关系时，求线性回归方程才是有意义的，才有助于根据 X_i 预测 Y_i。而像图 9-1a 和 9-1c 所示的数据，(X_i, Y_i) 间的关系或者是随机的，或者具有特殊的数据结构，虽然也可以按照公式(9-3) 和(9-4) 得到线性回归方程，但它并不能反映 X_i 与 Y_i 间的真实关系，因而也无助于对 Y_i 的预测。因此在使用回归方程前，首先要作假设检验，需要否定"X_i 与 Y_i 无关"的原假设，或者说需要制定一个判别标准。

下面用方差分析方法来帮助建立这个标准，为此需要引入和讨论回归平方和、相关系数等概念。可以证明对于任何一组数据 $(X_i, Y_i)(i = 1\cdots n)$，其总离差平方和总是等于残差平方和与回归平方和之和，即

$$\sum(Y_i - \overline{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \overline{Y})^2 \tag{9-6}$$

或

$$TSS = RSS + RSSR \tag{9-7}$$

公式(9-6)右边的第二项称为回归平方和。这个公式的含义是，回归分析只能解释总离差平方和中的一部分，即解释由于自变量 X 变化所引起的 Y 的变化，这一部分就是回归平方和。残差平方和是总离差平方和中的另一部分，是回归分析所不能解释的那一部分。残差平方和与回归平方和也都是统计量，而且可以证明，RSS 和 $RSSR$ 分别服从自由度为 $(n-2)$ 和 1 的 χ^2 分布。设想一种理想的情况，n 个实验点原来就在一条直线上，如图 9-4 所示，那么这条直线本身就是回归直线，回归平方和就等于总离差平方和，而残差平方和为零。可见残差平方和反映了实验数据偏离回归直线的程度。回归直线的参数 a 和 b 就是在要求残差平方和最小的条件下确定的。

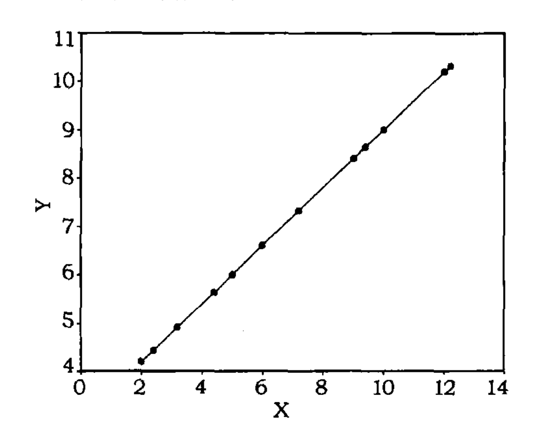


图 9-4 实验数据本身组成一条直线的一种理想情况，回归直线与实验直线重合

公式(9-6)中的每一项都是平方项，它们都是正值，因此残差平方和的取值范围总是处在 0 和总离差平方和 $\sum(Y_i - \overline{Y})^2$ 之间。现定义一个新的统计量 r^2：

$$r^2 = 1 - \frac{\sum(Y_i - \hat{Y})^2}{\sum(Y_i - \overline{Y})^2} = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} \tag{9-8}$$

r^2 反映回归分析所能解释总离差平方和的比例,它的取值在1和0之间。如果实验数据本身组成一条直线的理想情况(图9-4),则 $r^2 = 1$,而当实验数据 X_i 与 Y_i 完全无关时,r^2 接近于零。r^2 的开方值 r 称为样本 X_i 与 Y_i 间的皮尔逊相关系数,它是在 -1 和 $+1$ 间变化,r 的绝对值越接近1,表示 X_i 与 Y_i 间的相关程度越高;当 r 接近0时,表示 X_i 与 Y_i 间没有明显的相关关系。所以 r^2 和 r 都是相关强度的度量。r 的符号反映了相关的方向,是正相关还是负相关。利用回归平方和的关系式和公式(9-3):

$$RSSR = \sum(\hat{Y}_i - \overline{Y})^2 = \sum(a + bX_i - a - b\overline{X})^2 = b^2\sum(X_i - \overline{X})^2 = \frac{L_{xy}^2}{L_{xx}} \tag{9-9}$$

可以推导得到关于 r 的表达式:

$$r = \frac{L_{xy}}{(L_{xx} \cdot L_{yy})^{\frac{1}{2}}} = \frac{\mathrm{cov}(x, y)}{\sigma_x \cdot \sigma_y} \tag{9-10}$$

对于上面讨论的关于A型彩陶分布的例子,根据公式(9-10)可以计算得到,各遗址离彩陶生产中心的距离 X_i 与遗址中A型彩陶片的相对含量 Y_i 这两个变量间的相关系数 $r = -0.757$。如果变量 X 和 Y 是标准化的,那么 $\sigma_x = \sigma_y = 1$,它们之间的相关系数 r 就等于它们的协方差 $\mathrm{cov}(X, Y)$,即 $r = \mathrm{cov}(X, Y)$。相关系数 r 与斜率 b 的符号是一致的,在本例中 r 与 b 都是负的,因此是负相关,因为随 X 的增长,Y 是下降的。

通过上面的讨论,对回归方程有效性的检验过程如下:

(1) 提出原假设 H_0:X 与 Y 完全无关;备择假设为 H_1:X 与 Y 相关。

(2) 计算统计量。这里可以使用两个统计量进行检验,它们分别是 t 和 F。

$$t = \sqrt{(n-2)}\,\frac{|r|}{\sqrt{1 - r^2}} \tag{9-11}$$

可以证明,这个统计量服从自由度为 $(n-2)$ 的 t 分布。

另一方 面因为 RSS 服从自由度为$(n-2)$的 χ^2 分布和 $RSSR$ 服从自由度为1的 χ^2 分布,因而统计量

$$F = \frac{RSSR}{RSS/(n-2)} \tag{9-12}$$

服从 $F(1, n-2)$ 分布。对上面A型彩陶片按遗址距离分布例子,经计算得到。$t = \sqrt{12-2}\,\frac{0.757}{\sqrt{1 - 0.757^2}} = 3.664$ 和 $F = \frac{4263}{318/10} = 13.41$。

(3) 利用其中任一个统计量都可以对"X_i 与 Y_i 完全无关"的原假设进行检验。

对于 $t = 3.664$,查自由度 $df = 10$ 的 t 表,显示相应的显著性水平 $\alpha = 0.004$。对于 $F = 13.41$,查自由度 $df_1 = 1$ 和 $df_2 = 10$ 的 F 表,显示相应的显著性水平也为 $\alpha = 0.004$。由此可见这两个检验是等价的,都是在 $\alpha = 0.004$ 的显著性水平上拒绝"X_i 与 Y_i 完全无关"的原假设;以高达99.6 % 置信度接受备择假设,认为"X_i 与 Y_i 是相关的"。也就是说,如果这12对(X_i, Y_i)数据来自 X 与 Y 完全无关的总体,那么只有0.4% 的概率出现 $t > 3.664$ 或者 $F > 13.41$ 的情况。对于所讨论的实例,判断"X_i 与 Y_i 是相关的"的含义是"A型

彩陶生产点附近的遗址中 A 型彩陶片的相对数量与遗址离生产中心的距离是相关的，相距越远，每立方米堆积物中 A 型彩陶片的数量越少”。

由公式(9-11)可以看到，当样本容量 n 很大时，即使相关程度不高，r 数值离 1 较远，t 值也会较大，从而有可能以一定的置信度作出“X_i 与 Y_i 是相关的”的判断。也就是说对于大样本，即使实际的相关关系很弱，也有可能被检验出来。因此我们在实际研究中应同时关注(1)以一定的置信度判断是否相关和(2)相关系数本身的大小这两个方面。这类情况在下一章中讨论两个名称变量间的关联时同样存在。

9.2.3 线性回归中残差的分析*

从前面的讨论中，我们看到回归分析只是解释样本的总离差平方和中的一部分，解释了因自 变量 X 变化所引起的 Y 的变化。对于前面所讨论彩陶的实例，因距离的远近不同而导致的各遗址每立方米堆积物中 A 型彩陶片的数量变化通过回归分析得到了解释。各遗址中 A 型彩陶片相对数量变化中没有得到解释的那部分就反映在残差中。本小节将对残差做分析。下面的表 9-2 除重复列出表 9-1 中的原始数据 X_i 和 Y_i 外，还列出了回归值 $\hat{Y}_i$ 和残差$(Y_i - \hat{Y}_i)$。

表 9-2 12 个遗址中 A 型彩陶残片的密度和离生产中心的距离的回归分析中的残差

遗址号	1	2	3	4	5	6	7	8	9	10	11	12
X_i	4.0	7.0	15.0	20.0	21.0	24.0	28.0	29.0	33.0	35.0	36.0	44.0
Y_i	95.0	84.0	89.0	67.0	42.0	66.0	38.0	8.0	40.0	35.0	56.0	38.0
回归值	88.8	83.9	70.8	62.6	61.0	56.0	49.5	47.8	41.3	38.0	36.4	23.2
残差	6.2	0.1	18.2	4.4	-19.0	10.0	-11.5	-39.8	-1.3	-3.0	19.6	14.8

残差也是一个随机变量，也可以像对待其他随机变量一样来研究它的分布规律。对残差的分析有时能帮助寻找隐藏在未能被解释的那部分总离差平方和背后的原因。在所研究的实例中，也许第 8 号遗址会引起我们的注意。它的残差的绝对值特别大，而且是负值，说明在该遗址发现的 A 型彩陶片的相对数量比回归值低很多。产生这种情况的原因，有可能是因为该遗址交通不便，也可能是该遗址自己也生产彩陶，因此对输入外来彩陶的要求低所致。真正的原因需要有另外的考古资料来帮助判断，但残差分析显示 8 号遗址是一个比较特殊的遗址，需要引起注意。

9.3 相关分析的应用实例

9.3.1 仰韶文化陶器上刻划符号出现频率的相关性研究

在陕西省的半坡和姜寨两个距今约 6000 年的仰韶文化遗址的陶器上，都曾发现一些刻划符号(简称刻符)，王志俊(1980)统计共计约 42 种 243 个符号。有一种意见认为这些刻符是后期文字的雏形。文字作为信息的载体，人们交流的工具，它应该为相当广泛的地区的人们所共同使用。因此如果某些字词在该地区的某一地点为常用字词，那么这些

字词在该地区的另外地点也应被经常使用。反过来那些不常用的字词符号在这一地区的不同地点也应是共同的。如果这两个仰韶文化遗址出现的刻符属文字的雏形，它们也应表现出这种性质，某些刻符在半坡是常用刻符，其出现的频率高，它们在姜寨的出现频率也应高。反之亦然。我们用 X_i 与 Y_i 分别表示第 i 种刻符在半坡和姜寨出现的频率。下表列出王志俊统计的两地刻符的种类和数量，以及我们依此计算出的相应频率。

表 9-3　半坡和姜寨遗址的陶器上各类刻画符号的数量和频率统计

	半坡			姜寨		
刻画符号	数量	频率 Xi %	频率 Xi%	数量	频率 Yi %	频率 Xi%
ǀ	65	57.52	不计	72	55.38	不计
ǁ	4	3.54	8.33	7	5.38	11.86
Y	1	0.88	2.08	1	0.77	1.69
↾	9	7.96	18.75	5	3.85	8.47
T	2	1.77	4.17	1	0.77	1.69
∧	1	0.88	2.08	2	1.54	3.39
×	4	3.54	8.33	4	3.08	6.78
+	3	2.65	6.25	3	2.31	5.08
卄	1	0.88	2.08	1	0.77	1.69
⩤	4	3.54	8.33	8	6.15	13.56
⌐	4	3.54	8.33	2	1.54	3.39
⊣	6	5.31	12.50	2	1.54	3.39
V	0	0	0.00	2	1.54	3.39
9 种刻符	9×1	9×0.88	9×2.08	0	9×0	9×0
20 种刻符	0	20×0	20×0	20×1	20×0.77	20×1.69
总计	113	100	100	130	100	100

利用上面表中的数据，公式(9-10)给出这 42 种刻符在两地出现的频率（X_i，Y_i）之间的相关系数 $r = 0.989$，相关性极高。如果考虑到刻符"ǀ"出现的频率太高，相关分析中占的比重太大。而且刻符"ǀ"有时与陶器上偶尔产生的划痕不易分清。因此把刻符"ǀ"舍弃，重新计算其他 41 种刻符出现的频率，再作相关分析，得到相关系数 $r = 0.67$。利用公式(9-11)，计算得到 $t = 5.64$。已知自由度为(41-2) = 39，查 t 表，得 $t_{0.001}(n = 39) = 3.56$，小于样本的计算值 $t = 5.64$。因此可以以极高的置信度（$\alpha < 0.001$）判断，这些刻符在两地的出现频率之间是高度相关的，半坡和姜寨的常用刻符和偶用刻符是基本相同的。因此两地刻符出现频率的相关分析支持这些刻符具有"字词的使用频率在该文字系统的地区的各地点间存在相关性"性质的观点。当然这里的讨论仅限于刻符具有文字某种性质，至于仰韶陶器上的刻符是否真为文字的雏形，不属于本书讨论的范围。

顺便提及，半坡和姜寨相距仅 30 公里，在离两地约 100 公里的另一个仰韶遗址——李家沟遗址的陶器上也发现了相似的刻画符号，相关分析表明这些刻符的出现频率在李家沟与半坡两地的之间也是相关的。

9.3.2　赤峰地区中美联合考古调查中对稀疏分布的陶片的相关性分析

在 8.4 节中我们曾介绍了赤峰中美联合考古调查项目的例子，该项目的研究者们认

为,282个采集陶片数少于5片的采集点(他们称为“小采集”)是有考古意义的,这些稀疏散布的各时期陶片不是近期农业活动所致。他们的依据是一元方差分析和子弹形图所显示的各时期小采集陶片比例数的差异。本节将介绍他们用相关分析来佐证这个观点。如果陶片稀疏散布的小采集点的存在是近代为肥田而堆肥和土壤搬运所致,那么小采集中辽代等晚期陶片的比重应当超过大采集的。对1633个陶片采集点的两个变量,即采集点各时期的总陶片数(X)和晚期陶片的比例数(Y)作相关分析,得到显著性水平$\alpha = 0.079$和相关系数$r = 0.043$。反映采集的大小和晚期陶片的比重间的相关强度是非常弱的($r = 0.043$)。虽然表现出一定的显著性水平($\alpha = 0.079$),但那是由于样本容量甚大($n = 1633$)所导致的结果,9.2.2小节中曾提到对于大样本,很弱的相关关系也能被检验出来。因此相关分析支持X和Y不相关的原假设,即不支持稀疏散布的陶片是近代农业活动所致的意见。

赤峰考古调查的研究者们又从另一个角度探讨了小采集点存在的原因。如果它们是近期农业活动的结果,那么在现代村落附近小采集点的密度应增大。也就是说对不同时期而言,小采集点在总采集点中所占百分比(X)与同时期在现代村落附近的采集点在总采集点中所占的百分比(Y)之间应有较显著的相关关系,但实际上,在排除了兴隆洼,赵宝沟和小河沿等采集到陶片数很少的时期后,对于从红山文化、夏家店上层和下层、战国至汉,到辽代各时期上面两个百分比值之间的相关系数为0.103,而显著性水平$\alpha = 0.870$(原研究者计算的是斯皮尔曼相关系数,用以替代了本章讨论的皮尔逊相关系数,我们将在第十一章介绍斯皮尔曼相关系数,但在所讨论的例子中这个替代不影响后面的推论),因此可以以相当高的置信度接受X和Y不相关的原假设,即不支持近代农业活动“生成”小采集点的可能。两方面作相关分析的结论是一致的,并符合8.4节一元方差分析比较各时期小采集点比例数所得的结论。

赤峰考古调查资料的研究显示了相关分析怎样帮助检验近代农业活动是否是导致小采集点形成的原因。顺便提及,除本小节和8.4节所介绍的内容外,赤峰考古调查的研究者们还对小采集点作了其他定量方法的分析与验证,其目的是探讨在考古调查中怎样正确定义“遗址”这个似乎“不言而喻”的概念,重申了反对以“遗物”来替代“遗址”的观点。当然关于这个在考古学研究中重要但又有争议的概念的详细讨论,是超出本书的范围和作者的能力的。

9.3.3 相关分析考古应用的其他实例简介

7.4.4节曾介绍吴十洲(2001)对两周时期386座墓葬中青铜容器的数量和组合关系的统计研究工作。除了7.4.4节中使用的一元方差分析方法外,吴还计算了两周各时期墓葬中青铜鼎数量和其他青铜容器数量间的相关系数。这些相关系数都是正值,处于$r = 0.41$与0.92间。说明各时期总体上墓葬中如果鼎的数量多,则其他青铜容器的数量也多,但是各期的相关强度有差别。由此吴十洲认为“鼎位居于两周墓葬随葬青铜容器的中心地位,应继续予以肯定”,但“鼎与鼎的配置关系并不是一贯的,确定的,随时代不同而出现不同的以鼎为主的青铜容器配置关系,或不以鼎为主的青铜容器配置关系”。除吴十洲的工作外,米同乐等(1998)对有胡铜戈进行了回归断代。他们测量或整理了晚商、

西周、春秋和战国等4个时代124件铜戈的援长、胡长和内长数据，根据这3个变量建立了一个三元二次的回归方程用以预测未知出处铜戈的年代。据报道预测结果基本正确。鉴于三元二次回归方程已超出本书的范围，我们不作详细介绍。米同乐等认为回归分析用于考古各类器物的断代有一定应用前景。

9.4　线性相关和线性回归分析中的一些问题

9.4.1　相关与回归分析的比较

相关分析和回归分析都是研究一对随机变量间的非确定性关系，本书仅考虑随机变量间的线性关系。但是相关分析和回归分析研究问题的角度是有所不同的。前者关注于两个变量间关系的密切程度，即相关程度的强弱、能否通过显著性检验等，这里两个变量是平等的。后者更注重两个变量间的因果关系，因而区分自变量和应变量，希望在建立线性回归方程后通过自变量来预测应变量。另一方面相关分析和回归分析又是紧密相连的，只有当一对随机变量高度相关时，建立回归方程进行预测才有意义，预测才较为精确可靠。一般情况下相关分析比回归分析在考古研究中得到更为广泛的应用。前面对仰韶的刻符、赤峰地区各时期陶片数量的分析以及两周墓葬青铜鼎和其他青铜容器的数量分析均属相关分析的例子，而关于遗址中A型彩陶残片的密度随遗址距离分布的研究和对有胡铜戈的断代则属回归分析。

9.4.2　相关和回归分析的应用条件

线性回归分析的应用有一些前提条件。它要求各 Y_i 相互独立，它们的平均值聚集在一条直线的左右，它们的方差一致性等，这里不作详细讨论。但在进行回归分析前要先观察实验数据的散点图，考虑是否适宜做线性回归。最理想的情况是数据点分布在一个拉长了的椭圆的范围中。对于图9-1a、9-1c等的分布情况是不宜作线性回归分析的。有时实验点 X_i 与 Y_i 间呈抛物线关系或指数函数关系，这种情况需要先做变量转换，然后对 X 与$\sqrt{Y}$ 或 X 与 $\log Y$ 作线性回归分析。

在相关分析和回归分析时，一定要特别注意散点图上有没有偏离群体甚远的特殊点，它们有可能显著改变回归直线的参数。图9-5是这样一个例子，右上角的S点是一个特殊点，其余的6个点组成负相关，而由于S点的参加，使得它们共同组成正相关。在分析仰韶刻符的例子中，舍弃了刻符“|”，也是因为它出现的频率比其他刻符的频率高出10多倍，属于特殊数据。很多情况下特殊数据的出现是测量错误或记录错误的结果，但有时特殊数据却反映某种特殊的，应引起注意的现象。总之对线性回归中的特殊数据应该认真检查，决定取舍。

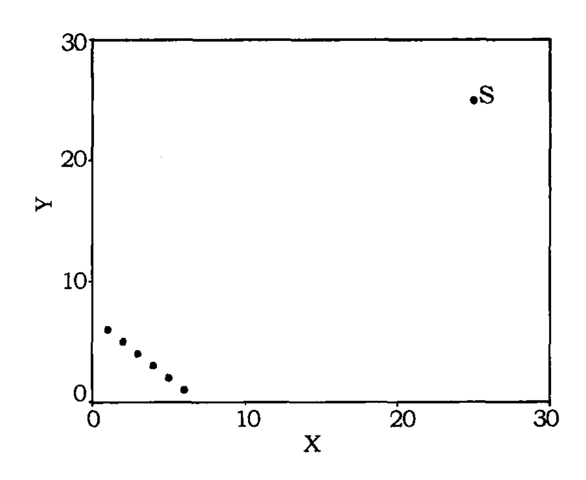


图 9-5　一组不适宜于回归分析的数据，特殊点 S 将显著改变回归参数

9.4.3　回归方程的稳定性和预测的误差*

9.2 节的讨论中见到回归分析中有残差存在。残差平方和 $\sum(Y_i - \hat{Y}_i)^2$ 服从自由度为 $(n-2)$ 的 χ^2 分布。残差平方和的平均值的开方被定义为剩余标准差或残余标准差：

$$s = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}} \tag{9-13}$$

剩余标准差 s 是一个很重要的统计量，它度量实验点相对于回归直线的"平均偏离"，它的数值决定回归直线的稳定性和预测的误差。

所谓回归直线的稳定性是这样理解的。设想在同一个总体中另外抽取 n 对数据 (X_i, Y_i)，利用新样本也可以建立一条新的回归直线 $\hat{Y}' = b'X + a'$。我们当然要关注两条回归直线的参数 a'，b' 与 a，b 之间的变动有多大，也就是关注回归直线的稳定性有多高。a 与 b 也是随机变量，它们的涨落用它们自己的标准差 s_a 和 s_b 来表述和度量。可以用下面的公式来计算 s_a 和 s_b：

$$s_a = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}} = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{L_{xx}}} \tag{9-14}$$

$$s_b = \frac{s}{\sqrt{L_{xx}}} \tag{9-15}$$

这两个公式表明，剩余标准差 s 直接决定了 s_a 和 s_b 的大小。而在 s 不变的情况下，实验点越多（n 大和内积系数 L_{xx} 大）和自变量的变动范围越大，则 s_a 和 s_b 越小，回归直线的稳定性也越高。

回归分析的目的之一是给定 X_0 来预测 Y_0，Y_0 的标准差 s_{Y_0} 由回归分析的残余标准差 s 和 a，b 的误差共同导致，其计算公式如下：

$$s_{Y_0} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{L_{xx}}} \tag{9-16}$$

可以看出，预测的精确度除与实验点的多少有关外，还与 X_0 的位置有关。当 X_0 接近平均值 $\bar{X}$ 时，预测的精确度最高，而当 X_0 不断偏离 $\bar{X}$ 时，预测的误差也不断增大。因此当回归直线确定后，内插的预测误差小，而外推的预测是相当危险的，可能会导致很显著的误差。

9.4.4　关于多元情况下的线性回归问题

本章仅考虑了一元的情况，即 Y 仅依赖于一个自变量 X。当 Y 依赖于若干个自变量（$X_1, X_2, X_3, \cdots$）时，情况就要复杂得多。当然在数学上我们可以不考虑其他自变量对 Y 的影响，而只考虑 Y 与其中一个自变量 X_i 的关系，求它们之间的相关系数。这个相关系数称为 Y 对某个自变量的简单相关系数。但简单相关系数已不像一元情况那样能确定地反映 Y 与 X_i 之间的真实关系，它可能受到其他自变量的影响。需要在控制其他变量的条件下求 Y 与各个自变量间的相关系数，称为偏相关系数。偏相关系数才比较真实地反映变量间的关系。此外也需要考虑自变量之间的关系。当然这些内容已超出本书的范围，这里仅是提醒读者注意这类问题，读者在有需要时可查看有关多元回归分析的书籍。

第十章　名称变量间关联的假设检验

第五章至第九章主要讨论数值型的变量，研究实体相对于数值变量的分布，比较样本所属总体间平均值是否有差别等。第九章又介绍了两个数值变量之间的相关问题。本章将主要讨论名称变量之间关联的研究。相关和关联都是涉及变量之间关系的研究，但因为变量的层次不同，所用的数学方法不同，也用了不同的名称。相关和关联分别是从英语的 correlation 和 association 两个词翻译过来的。本章的最后一节将讨论实体按单个名称变量分布的假设检验。

10.1　2×2 四格交叉列联表的 χ^2 检验

10.1.1　名称变量间关联 χ^2 检验的原理和过程

首先讨论名称变量中最简单的情况，即二元变量间的关联问题。第八章 8.3 节曾根据两个墓地中带或不带随葬品的墓葬数目的统计，检验总体上两墓地带随葬品墓的比例数间是否有差别。8.3 节中的数据列表表示如下：

表 10-1　两个墓地中带或不带随葬品的墓葬数的统计表

	带随葬品	不带随葬品
甲地的墓葬数	60	40
乙地的墓葬数	35	15

这种类型的表格称为“2×2 列联表”，也称四格表。表的第一列和第一行分别表示两个名称变量的名称和它们的两个取值状态，即墓葬的所在地和是否带随葬品。表格的主体是右下部的 2 行 2 列共 4 格，记录了甲乙两地、带或不带随葬品 4 种交叉状态的实体数目，因此列联表又称为实体交叉分类频次表。8.3 节是检验两个墓地中带随葬品的墓葬的比例数是否一致。比例数是数值型变量，因此 8.3 节进行的是参数的假设检验。但也可以从另外一个角度提出问题，将全部 150 座墓葬看成从一个单一总体中抽取的样本，“墓葬所在地”和“是否带随葬品”是描述每个墓葬的两个属性。当然这是两个二元的名称属性，每个属性只有两个被允许的取值。现在要根据这个样本来检验，对于总体而言“墓葬所在地”与“墓葬是否带随葬品”这两个变量间是否有关联。在回答这个问题前，我们先把表 10-1 改写如下：

表 10-2　两墓地中带或不带随葬品的墓葬数统计表的改写

	带	不带	和
甲地	60	40	100
乙地	35	15	50
和	95	55	150

表 10-2 比表 10-1 添加了 1 行 1 列。在添加的最后一行(列)列出每列(行)元素的总和,也称列(行)变量的边缘和或边缘分布。甲(乙)墓地的墓葬总数为 100(50),带(不带)随葬品的墓葬总数为 95(55)。

名称变量间关联问题是用统计量 χ^2 来检验的。检验的步骤与前几章数值变量的假设检验是相似的。检验过程如下:

(1) 第一步先提出原假设 H_0:假设"墓葬是否带随葬品"与"墓葬所在地"两个变量间没有关联;相应的备择假设 H_1 为:这两个变量间存在关联。

(2) 第二步是在原假设 H_0 成立的前提下,计算两个变量交叉取值的 4 种状态的墓葬数的期望值。先计算墓地甲带随葬品的墓葬数的期望值,它应该等于"带随葬品墓的总数"和"墓地甲的墓葬数在总墓葬数中所占的比例"之乘积,即 $95 \times \frac{100}{100+50} = 63.3$。而墓地甲不带随葬品墓葬数的期望值应为 $55 \times \frac{100}{100+50} = 36.7$。同样可以计算墓地乙带和不带随葬品等两种情况墓葬数的期望值。这 4 个期望值用 E_i 表示,列入表 10-3 中。

表 10-3　"无关联"假设前提下墓葬交叉分类的期望频次表

	带	不带	和
甲	63.3	36.7	100
乙	31.7	18.3	50
和	95	55	150

用 Q_i 表示四格表 10-2 中第 i 格的实际观测值。可以证明统计量:

$$\chi^2 = \sum \frac{(Q_i - E_i)^2}{E_i} \tag{10-1}$$

服从 χ^2 分布,其自由度为:

$$df = (\text{行数} - 1) \times (\text{列数} - 1) \tag{10-2}$$

对于我们的例子:

$$\chi^2 = \frac{(60-63.3)^2}{63.3} + \frac{(40-36.7)^2}{36.7} + \frac{(35-31.7)^2}{31.7} + \frac{(15-18.3)^2}{18.3} = 1.435$$

四格表的自由度为$(2-1)\times(2-1)=1$。自由度是等于能自由赋值的单元格的数目。在所研究的例子中,两地的墓葬总数和带或不带随葬品的墓葬数都是固定的,因此 4 个单元格中只允许对一个单元格自由赋值。一个单元格赋值后,其他 3 个单元格的值就自动被确定了,因此自由度等于 1。

(3) 第三步进行判断,查自由度为 1 的 χ^2 表,对应于 $\chi^2 = 1.435$ 的显著性水平是 0.23。这个 χ^2 检验的意义是:对于一个假想的,墓葬所在地和是否带随葬品完全无关的总体,随机抽取 150 座墓葬,那么有 23% 的概率抽样到一个如表 10-1 所示那样偏离期望值或偏离更大的样本。因此在显著性 $\alpha = 0.2$ 的水平上不能拒绝"墓葬是否带随葬品与墓地间无关联"的假设。回忆第八章的 8.3 节曾对表 10-1 所示的样本,用二项式分布检验两类墓葬的比例数是否有显著差别,计算了统计量 Z,得到 $Z = 1.2$,查正态分布表得到的显著性水平 α 也是 0.23,接受了两地带随葬品墓葬的比例数无差别的原假设。两种检验的

角度不一样,方法不一样,但检验的结果是一致的。

对于 2×2 的四格表,也可以用另一种较简便的方法计算 χ^2 值。把表 10-1 写成一般形式:

表 10-4 四格表的一般形式

	Y(+)	N(-)
A(+)	a	b
B(-)	c	d

χ^2 也可由下面的公式计算得到:

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(b+d)(a+c)(c+d)} \tag{10-3}$$

式中的 $n = (a+b+c+d)$,代表总的实体观测数。对于四格表,公式(10-1)和(10-3)是等效的。对上述的实例验算如下,将 $a = 60$, $b = 40$, $c = 35$ 和 $d = 15$ 代入式(10-3),得到:

$$\chi^2 = \frac{150(60 \times 15 - 40 \times 35)}{(60+40)(40+15)(60+35)(35+15)} = 1.435$$

可见两个公式计算的结果是一致的。

10.1.2 样品的容量对 χ^2 检验的影响

本小节将讨论四格表 χ^2 检验中的两个重要的问题。(1) 四格表 χ^2 检验的结论仅仅涉及两个名称变量之间是否有关联,并不能给出关联的强度有多大。(2) 当 4 个单元格的数值按相同的比例增加时,χ^2 值也将按同样的比例增加。这从公式(10-3)可以清楚地看出,因为分子的因次是单元格数值的 5 次方,而分母只是 4 次方。四格表的自由度是不变的,由此 χ^2 值的增大会使检验的显著性升高,并有可能会改变假设检验的结论。下面我们把表 10-2 中每个单元格的频次值增加 4 倍,总的墓葬数也从表 10-2 中的 150 座增加为 600 座。对于扩大了容量的样本,其四格表如下:

表 10-5

	带随葬品	不带随葬品	和
甲墓地	240	160	400
乙墓地	140	60	200
和	380	220	600

将表 10-5 的数据代入公式(10-3),计算得到 $\chi^2 = 5.741$,正好是原来对表 10-2 计算的 $\chi^2 = 1.435$ 的 4 倍。对表 10-5 数据作 χ^2 检验,自由度仍是 1,查 χ^2 表,相应的显著性水平 $\alpha = 0.0166$。因此可以以 98.3% 的置信度拒绝"墓葬是否带随葬品与墓地没有关联"的原假设,而接受这两个变量之间是关联的备择假设。这看上去似乎与 10.1.1 中的检验结论是矛盾的。两个样本带随葬品墓葬数的百分比是相等的,都等于:

甲墓地带随葬品墓葬数的百分比 $\frac{60}{100} = \frac{240}{400} = 60\%$

乙墓地带随葬品墓葬数的百分比　$\frac{35}{50}=\frac{140}{200}=70\%$

两墓地带随葬品墓葬数的平均百分比　$\frac{95}{150}=\frac{380}{600}=63.3\%$

但根据这两个样本对总体的“墓葬是否带随葬品与墓地间是否有关联”的判断却是不一致的。小容量样本所未能检验出的关联,在大容量样本的情况下却被检验发现,即大样本对于发现弱的关联更敏感。在第八章讨论总体比例数的检验时,也曾专门讨论了样本的容量对检验结论的影响。为了显示这个影响,下面按 8.3 节中正态分布 U 检验的方法,根据表(10-5)的数据,对甲乙两墓地带随葬品墓葬的总体比例数是否一致作检验,利用公式(8-3),求得 $s_{Dp}=\sqrt{0.633\times0.367\times\frac{400+200}{400\times200}}=0.0418$,从而,$Z=\frac{|(\hat{p}_1-\hat{p}_2)-(p_1-p_2)|}{s_{Dp}}=\frac{|(0.6-0.7)-0|}{0.0418}=2.40$。查正态函数表,显著性水平 $\alpha=0.0166$。因而同样以 98.3 % 的置信度拒绝“两墓地带随葬品墓葬的比例数无显著差别”的原假设。回忆在 8.3 中对表(10-2)的数据,墓葬总数为 150 时,也曾经在 $\alpha=0.23$ 的显著性水平上接受“两墓地带随葬品墓葬的比例数无显著差别”的假设。因此无论对于 χ^2 检验,或者对于利用正态分布的 U 检验,都是当样本容量增加时,对于发现弱的关联更敏感,关联更易被检出。在第九章利用公式(9-11)作两个数值变量相关性的显著性检验时,也同样观察到样本的容量对假设检验的影响。因此第九章曾强调要“同时关注相关性检验的置信度和相关系数 r 本身的大小两个方面”,相关系数反映了相关关系的强弱。在讨论名称变量之间的关联时,除注意 χ^2 检验给出的接受或拒绝关联假设的置信度外,也应该寻找一个相应的统计量来反映关联的强弱。这里同样需要同时关注接受或拒绝关联的置信度和关联本身的强度。以相当高的置信度检验出很弱的关联往往是没有实际意义的。

10.1.3　名称变量间关联强弱的度量

表征名称变量间关联强弱的统计量有 ϕ 系数和 Yule Q 系数。现分别予以介绍。

(一) ϕ 系数。

为了寻找反映关联强弱的统计量,考虑到 χ^2 值是正比于样本的容量 n 的,很自然会想到用

$$\phi=\sqrt{\frac{\chi^2}{n}} \tag{10-4}$$

作为关联强弱的度量,这样定义的 ϕ 系数是与样本的容量 n 无关的。利用公式(10-2),得到

$$\phi=\sqrt{\frac{\chi^2}{n}}=\sqrt{\frac{(ad-bc)^2}{(a+b)(b+d)(a+c)(c+d)}}=\frac{|ad-bc|}{\sqrt{(a+b)(b+d)(a+c)(c+d)}} \tag{10-5}$$

希腊字母 ϕ 读作“phi”。对于两个二元变量的四格表。ϕ 是在 0 与 1 之间变动,$\phi=1$ 表示两个名称变量间的完全关联,$\phi=0$ 表示两个变量间不存在关联。对于表 10-2 或表 10-5

所列的样本数据,计算得到它们的 ϕ 值分别为$\sqrt{\frac{1.435}{150}} = 0.098$ 和$\sqrt{\frac{5.741}{600}} = 0.098$,即两组数据的关联强度是相等的,而且关联是很弱的。在四格表中 a 代表两属性联合取正值的实体数,d 代表两属性联合取负值的实体数,因此 $a \times d$ 反映两个属性间的正协变,而 $c \times d$ 反映两个属性间的负协变。公式(10-5)中的分子是正、负协变的差值。

(二) Yule's Q。

除 ϕ 外,也常用另一个称为 Yule's Q 的系数来表示四格表关联的强弱,Q 是这样定义的:

$$Q = \frac{ad - bc}{ad + bc} \tag{10-6}$$

公式(10-6)中分式的分子也是正、负协变的差,而其分母为正、负协变的和。Q 的绝对值与 ϕ 相似,也是在0与1之间变动,但 Q 可正可负,相应表示正关联和负关联。同样,$Q = \pm 1$ 表示两个名称变量间的完全关联,$Q = 0$ 表示两个变量间不存在关联。在下一章的11.2节可以看到 Q 系数是 *Gamma* 等级相关系数的一种特殊情况。对于表10-2或表10-5的数据,利用公式(10-6)计算得到,Q 都是等于$\frac{60 \times 15 - 40 \times 35}{60 \times 15 + 40 \times 35} = 0.217$,同样说明关联是较弱的。$\phi$ 与 Q 也有一些不同之处,一般来说 $|Q| > \phi$,因此更易观察到弱的关联。另外对 ϕ 而言,必须 a 和 d 均为0,或 b 和 c 均为0,ϕ 才等于1,才判断为完全关联。但是对于 Q,只要 a、b、c,或 d 中的任意一个为0,Q 就等于1,判断为完全关联。一般情况下可随意选取 ϕ 或 Q 度量关联强度,但当有的单元格中的频次数很低时,需要根据所研究的实际问题考虑选哪个系数更合适。

在本小节的最后,再次强调,对于四格列联表检验结果的表述,应同时说明检验的置信度和关联的强度。例如对于四格表10-2和10-5关于墓地与墓葬是否带随葬品间关联的检验结果应分别写为($\chi^2 = 1.435, \alpha = 0.23, \phi = 0.098$)和($\chi^2 = 5.741, \alpha = 0.017, \phi = 0.098$),或者写为($\chi^2 = 1.435, \alpha = 0.23, Q = 0.217$)和($\chi^2 = 5.741, \alpha = 0.017, Q = 0.217$)

10.1.4 四格表 χ^2 检验的前提条件

前面几章讨论数值变量的各种假设检验中,曾十分注意每种检验方法的假设前提,例如要求样本服从正态分布,要求样本间的方差一致,要求不存在偏离群体太大的特殊数据点等。在列联表的检验中涉及的是名称变量,进入单元格中的数据只能是频次,这里所要求的前提条件是样本的容量要足够大,而且要求每个单元格的期望值 E_i 不应太小。当 E_i 值太小时,由于随机涨落,使得 χ^2 值波动大而判断失误。至于具体要求 E_i 的最小数值多大,在各统计学的书中并没有统一的规定。一般要求所有的 E_i 值不小于5,应该是比较保险的,有的情况下可放宽条件,例如要求 E_i 值小于5的单元格的比例低于20%。如果在实际研究中出现过多和过小的 E_i 或 Q_i,可以用费舍公式来精确计算相应的概率值 P:

$$P = \frac{(a+b)!(a+c)!(b+c)!(b+d)!}{n!a!b!c!d!} \tag{10-7}$$

在计算机普及以前,费舍公式(10-7)的计算量可能是令人头疼的,但是有计算机的帮助这

已是很简单的工作了。对于四格表 10-2 和 10-5,公式(10-7)分别给出 $P = 0.28$ 和 $P = 0.019$,稍高于 χ^2 检验的 $\alpha = 0.231$ 和 $\alpha = 0.017$。

10.1.5 关于 χ^2 检验中的连续性修正

在有的统计书和统计软件中,考虑到 2×2 列联表的格数太少,为减少反映频次的离散型变量 和连续型变量 χ^2 间的差异,对计算 χ^2 的公式(10-1) 要作所谓的连续性修正。修正的公式如下:

$$\chi^2 = \sum \frac{(|Q_i - E_i| - 0.5)^2}{E_i} \tag{10-8}$$

对于四格表 10-2 作连续性修正后的 χ^2 值为 1.037 (未修正的值为 1.435),对应的显著性水平 $\alpha = 0.309$ (未作修正时为 $\alpha = 0.231$)。连续性修正一定程度上减少了因随机涨落引起的 χ^2 值的偏大。对于四格表 10-5,因为样本容量大了 4 倍,连续性修正所导致的相对改变就要小得多,修正后的 χ^2 值为 5.319 (未修正的值为 5.742),α 值由未修正时的 0.017 改变为 0.021。连续性修正主要应用于样本容量小,E_i 和 Q_i 值较低的四格表的检验。

10.2 四格表的关联检验中第三变量的引入和因果关系考察中的复杂性

在两个二元名称变量间的关联检验中。如果在变量间能区分出自变量和应变量,这就有可能进一步作因果关系的考察。但是必须十分小心,因为有时候表现出来的关联带有"假象"的成分,关联的背后可能有第三个变量在起作用。下面分析两个具体例子(例子均引自 S.Shennan 的《Quantifying Archaeology》)。

例一 表 10-6 统计了 128 座墓葬,并按墓主人的性别和墓穴的大小分类。这里可以把墓主人的性别看成自变量,考察墓穴的尺寸是否依赖于墓主人的性别,因此后者是应变量。

表 10-6 128 座墓葬按墓主人性别和墓穴大小的分类表

	小墓穴 Y	大墓穴 not Y
男 X	22	47
女 not X	33	26

检验结果为($\chi^2 = 7.505, \alpha = 0.006, \phi = 0.242, Q = -0.461$),可以以 99.4% 的置信度判断墓穴的大小是与墓主人的性别关联的,而且关联系数 Q 也不太小,达 0.461,说明关联并非太弱。负的 Q 值表示男性墓葬的墓穴大的比例高。这里说明一下,如果把两列或两行数据交换一下,Q 就是正值了,因此关联的正负号的情况是与变量的取值在表中怎样排列有关的。我们把上面的 Q 值称为 X 与 Y 间的零极关联系数。

上面检验的结论是墓主人性别与墓穴大小间有一定的关联,因为把性别看成自变量,则似可进一步导出优葬男性和女性地位低的考古学推论。但这样推论是危险的。如果引入墓主人身高这个因素,同时考虑性别、墓穴大小和墓主人身高等 3 个变量,问题就

复杂化了。假设测量了这128具人骨的高度，以155cm为界，把人骨分成高和矮两类，这样样本需按三个变量进行分类，现将分类结果总结于表10-7中。

表10-7 128座墓葬按墓主人的性别和墓穴大小分类表，以墓主人身高为控制变量

		小墓穴 Y	大墓穴 not Y
矮（t）	男 X	17(a)	4 (b)
矮（t）	女 not X	29 (c)	6 (d)
高（not t）	男 X	5 (e)	43 (f)
高（not t）	女 not X	4 (g)	20 (h)

表(10-7)实际上是上下两个四格表，分别以身高的高矮为参数。现对高矮两类人骨分别检验性别与墓穴大小的关系。结果如下：

对矮的人骨：($\chi^2 = 0.032, \alpha = 0.857, \phi = 0.024, Q_{xyt} = -0.064$)

对高的人骨：($\chi^2 = 0.57, \alpha = 0.450, \phi = 0.089, Q_{xy\bar{t}} = -0.26$)

上面 Q 的下标 xyt 和 $xy\bar{t}$ 分别表示在 t 的两个不同取值条件下计算的 xy 间的 Q 值。很明显当把人骨分成高矮两类后再分别检验性别与墓穴大小的关系时，诸 χ^2 值和 Q 值均很小，墓主人性别与墓穴大小间不存在关联。因此前面不区分人骨的高矮，根据性别与墓穴大小分类计算的 χ^2 值和零极关联系数 Q 值作出的优葬男性的推论是错误的。当有三个或更多的变量存在，而且我们认识到它们之间可能有复杂的关系时，就不能像10.1节中那样简单地仅考虑一对变量之间的零级关联系数，而必须同时考虑第三个变量对另两个变量间零级关联系数的影响，并在控制第三变量条件下计算另两个变量间的一级关联系数。例如对于这批墓葬，当我们得到性别和墓穴大小可能有关联的推论后，而又怀疑人骨的高矮对墓穴的大小也可能有影响时，应该在控制第三变量（人骨高矮）条件下计算性别和墓穴大小间的一级关联系数 $Q_{xy,tied-t}$，下标“$tied-t$”表示受控于变量 t。计算一级关联系数的公式是：

$$Q_{xy,tied-t} = \frac{(ad + eh) - (bc + fg)}{(ad + eh) + (bc + fg)} \qquad (10\text{-}9)$$

对于上面分析的实例 $Q_{xy,tied-t} = \dfrac{(17 \times 6 + 5 \times 20) - (4 \times 29 + 43 \times 4)}{(17 \times 6 + 5 \times 20) + (4 \times 29 + 43 \times 4)} = -0.175$。$Q_{xy,tied-t} = -0.175$ 在数值上显著小于未受人骨高矮控制的零级关联系数 $Q_{xy} = -0.461$。控制人骨高矮后，性别和墓穴大小间的关联变弱，说明10.1节中关于性别和墓穴大小间的零级关联中是有相当的虚假成分的。

可以进一步考察人骨高矮(t)和墓穴大小间(y)的关系。重新整理表10-7，得到表10-8。

表10-8 128座墓葬按墓主人身高和墓穴大小分类表，以墓主人性别为控制变量

		小墓穴 Y	大墓穴 not Y
男 X	矮（t）	17	4
男 X	高（not t）	5	43
女 not X	矮（t）	29	6
女 not X	高（not t）	4	20

计算人骨高矮和墓穴大小的零级关联,得(χ^2 = 62.3,α = 0.000,ϕ = 0.698,Q_{xy} = 0.940)。再计算控制墓主人性别条件下人骨高矮和墓穴大小的一级关联系数 $Q_{xy,tied-x} = \frac{(17\times43+29\times20)-(4\times5+6\times4)}{(17\times43+29\times20)+(4\times5+6\times4)} = 0.935$。可见人骨高矮和墓穴大小间的零级和一级关联系数 Q 都非常大,而且第三变量——墓主人性别的引入对关联系数的大小改变不很大。同时注意到 χ^2 值很大,由此可以以极高的置信度判断,人骨高矮和墓穴大小两个变量间存在非常强的关联,而且它们之间的关联基本上不受第三变量 — 墓主人性别的影响。因而可以进一步作因果关系的判断:人骨的高矮决定了墓穴的大小。现在更清楚地看到,当不考虑人骨高矮时,性别与墓穴大小所表现出来一定程度的关联,实际上是因为性别与人骨高矮间有关联所导致,男性人骨中高的比例大。性别与墓穴大小之间的关联仅是表观的,不是实质的。

例二　表 10-9 统计了 212 座墓葬,并按年代早晚(t)、墓主人的性别(X)和随葬是否有手镯(Y)3 个变量分组。

表 10-9　212 座墓葬按墓主人性别和是否随葬手镯分类,以墓葬分期为控制变量

		有手镯(Y)	无手镯(非 Y)
早期(t)	男(X)	31	27
早期(t)	女(非 X)	25	5
晚期(非 t)	男(X)	11	39
晚期(非 t)	女(非 X)	28	36

希望考察墓主人性别和墓葬中带不带手镯之间是否有关联。为此分 4 种情况计算 X 与 Y 间的 Q 值。

全部墓葬不分早晚期　$Q_{xy} = -0.34$

早期墓葬　$Q_{xyt} = -0.62$

晚期墓葬　$Q_{xy-not-t} = -0.47$

控制墓葬分期　$Q_{xy-tied-t} = -0.524$

比较这 4 个 Q 值可看到,对早、晚期的墓,墓主人的性别和墓葬是否带手镯间的关联均强于不分期混合计算的关联强度,控制第三变量时代分期 t 后,全部墓葬的性别与是否带手镯间的一级关联系数在数值上大于零极关联系数。这说明不考虑年代早晚因素时,这个关联某种程度上被隐藏了。为什么会发生这种情况呢,因为对于早、晚两期的墓葬,性别与是否随葬手镯的关系是不同的,早期很少有妇女不带手镯,而晚期很少有男子带手镯。早晚期不同的数据结构要求分期考察墓主人的性别与带手镯的关联,或者在控制年代分期的条件下考虑墓葬的性别与带手镯的关联。

前面两个例子提醒我们,当考察两个名称变量 X 与 Y 间的关联时,必须根据考古学的知识分析是否有别的因素可能影响这两个变量间的关系。第三变量可能增强,也可能减弱 X 与 Y 间表观的关联,还可能揭示出 X 与 Y 间存在更复杂的关系。X,Y 与 t 三个变量可以计算 12 个 Q 值,如果有 4 个变量,可能组成的 Q 值数目就更多,计算工作量更大。当然没有必要计算全部 Q 值,关键是依据具体的研究目的和我们已掌握的考古学知识来判

断,需要考察哪一对变量间的关联,是否需要控制别的变量。

10.3 $r \times c$ 列联表的 χ^2 检验和关联强度系数 V

前面讨论二元名称变量间的关联,二元变量只能有两个状态,只能取两个值,因此两个二元变量列联时,得到一张四格表。一般名称变量可以有多个状态,如果列变量 X 有 r 个状态,行变量 Y 有 c 个状态,则观测数据将组成一张 $r \times c$ 的列联表(见表 10-10)。第 i 列第 j 行的单元格中记录实体取值为第 i 个列变量值和第 j 个行变量值的频次数 n_{ij},这是列联表的主体。经常在表的最后增加一行(列),记录每列(行)全部元素的和,称为列(行)变量的边缘和或边缘分布。n 为样本中全部实体的总数。

表 10-10 $r \times c$ 列联表

	X_1	X_2	…	X_r	行和
Y_1	n_{11}	n_{21}	…	n_{r1}	n_{*1}
Y_2	n_{12}	n_{22}	…	n_{r2}	n_{*2}
…	…	…	…	…	
Y_c	n_{1c}	n_{2c}	…	n_{rc}	n_{*c}
列和	n_{1*}	n_{2*}	…	n_{r*}	n

下面通过两个实例来讨论多状态名称变量之间的关联问题。

例一 表 10-11a 统计了某墓地 136 座墓,并按墓式和墓主人的年龄段分组。墓式有土坑、木制墓室和石砌墓室等 3 类,年龄段也分成 3 段。因此得到一个 3×3 的表,共有 9 个单元格,每个单元格中记录了相应墓式和年龄段的墓葬数目 n_{ij}。表的最后一列(行)记录同类墓式(同一年龄段)墓葬的频次和 n_{*j} 和 n_{i*}。需要检验墓式与墓主人年龄段间是否有关联。所用的方法和处理四格表的情况是相同的,也是作 χ^2 检验。

表 10-11a 136 座墓葬按年龄段和墓室结构分类的实际观测频次表

墓式 \ 年龄段	青少年	壮年	老年	行总和 n_{*j}
简单土坑	23	19	11	53
木制墓室	12	17	13	42
石砌墓室	10	16	15	41
列总和 n_{i*}	45	52	39	136

检验过程如下:

(1) 作原假设 H_0:“墓式与墓主人年龄段之间不存在关联”。

(2) 在这个假设前提下,计算每个单元格的期望频次值 E_{ij},计算方法如公式(10-10)所示。

$$E_{ij} = \frac{n_{i*} \times n_{*j}}{n} \tag{10-10}$$

例如,(青少年,简单土坑)的期望频次值应为 $\frac{(45 \times 53)}{136} = 17.54$,其他 8 个期望频次值也类似计算。将计算结果写入表 10-11b。表 10-11b 和表 10-11a 的边缘分布是一致的。

表 10-11b　无关联假设前提下 136 座墓葬按年龄段和墓室结构分类的期望频次表

墓式＼年龄段	青少年	壮年	老年	行总和
简单土坑	17.54	20.26	15.20	53
木制墓室	13.90	16.06	12.04	42
石砌墓室	13.57	15.68	11.76	41
列总和	45	52	39	136

(3) 计算统计量：

$$\chi^2 = \sum_i \sum_j \frac{(E_{ij} - n_{ij})^2}{E_{ij}} \tag{10-11}$$

这个统计量服从自由度为

$$df = (r - 1)(c - 1) \tag{10-12}$$

的 χ^2 分布。对于所分析实例，公式(10-11)给出 $\chi^2 = 5.169$，自由度等于$(3-1)\times(3-1) = 4$。

(4) 查 χ^2 表，得到 $\alpha = 0.27$。因此在 27 % 的显著性水平上，接受原假设：没有观察到墓式与墓主人年龄段之间有明显的关联。

前面在讨论变量间相关和关联中，我们一直强调，仅仅以一定的置信度接受或拒绝关于变量间关联的假设是不够的，必须同时考察相关强度或关联强度。对于二元变量的 2×2 四格表，曾定义了 ϕ 和 Q 两个量来度量关联的强度。在分析 $r\times c$ 列联表中变量间的关联时，也必须注意关联强度。因此也需要定义相应的关联强度系数。$r\times c$ 列联表单元格的数目超过4，无法计算 Q 值。虽然仍可以计算其 ϕ 值，但 ϕ 值已不局限于(0—1)之间，它可以大于 1，且随行数和列数的增加而发散。因此 ϕ 和 Q 不能作为多状态名称变量间关联强度的指标，需要另外定义一个量，称为 Cramer's V，作为关联强度的度量。

$$V^2 = \frac{\phi^2}{\min[(r-1),(c-1)]} \tag{10-13}$$

公式中的分母表示从 $(r-1)$ 和 $(c-1)$ 中选择数值小的数字。式(10-13)定义的 V，其数值是限定于 1 与 0 之间的，V 的数值越大，反映关联强度越强。对于本节的实例，$\phi^2 = \frac{\chi^2}{n} = \frac{5.169}{136} = 0.038$，$V = \frac{\phi}{\sqrt{(3-1)}} = 0.138$。$V$ 值接近 0 而离 1 很远，因此说明关联是很弱的，这与前面 χ^2 检验中接受“未观察到显著关联”的原假设是符合的。

10.4　用预测中误差降低的比例来度量变量间的关联，λ 与 τ 系数*

第九章在讨论数值型变量间的相关时，曾指出回归方程的建立可以降低预测应变量时的误差，相关程度愈高，预测的误差愈小。这种情况同样适用于名称变量。举例来说，有一个人群，按编号预测每一个人的性别，误差可能会比较大。但是如果每个人的职业是已知的，再预测这个人群中每个人的性别，误差就会变小。因为职业与性别这两个名称变量之间是有一定程度的关联的，例如医院的护理人员女性占多数，而出租车司机男

性占多数。现在作一个逆向的思考,能否设想将"在知道了自变量的取值后,对预测应变量取值误差的降低程度"作为它们之间关联程度的度量呢?这种度量称为 *PRE* 度量,取自英语 Percentage of Reduced Error 三个字的第一个字母。*PRE* 是这样定义的

$$PRE = \frac{E_1 - E_2}{E_1} \tag{10-14}$$

式中 E_1 表示未知 X 与 Y 的关系时,预测应变量 Y 的误差,而 E_2 表示已知 Y 与 X 的关系时,利用已知的关系预测应变量 Y 的误差。$(E_1 - E_2)$ 反映利用已知的关系进行预测时误差的减少。因此 *PRE* 值就是利用已知的关系进行预测时误差减少的比例。*PRE* 与 Cramer's V 相似,其变化范围也是0与1间。如果 Y 与 X 间不存在相关或关联,则 $E_1 = E_2$,因此 $PRE = 0$。另一个极端,如果 Y 与 X 间完全关联,即它们间存在着函数关系,从而 X 能完全地确定 Y 的取值,这样 E_2 就等于0,$PRE = 1$。有两种不同的计算 E_1 与 E_2 方法,相应有2个不同的 *PRE* 系数,分别是 λ 系数和 Goodman and Kruskal 的 τ 系数。需要指出 *PRE* 度量适用于各种层次的变量,并不限于名称变量。

10.4.1 *PRE* 的 λ 系数

为了后面的讨论方便,这里重新抄录 $r \times c$ 列联表 10-10。

表 10-10 $r \times c$ 列联表

	X_1	X_2	…	X_r	行和
Y_1	n_{11}	n_{21}	…	n_{r1}	n_{*1}
Y_2	n_{12}	n_{22}	…	n_{r2}	n_{*2}
…	…	…	…	…	
Y_c	n_{1c}	n_{2c}	…	n_{rc}	n_{*c}
列和	n_{1*}	n_{2*}	…	n_{r*}	n

因为涉及预测问题,需要区分自变量 X 和应变量 Y。表中列变量是自变量 X,有 r 个状态,行变量是应变量 Y,有 c 个状态。i 列 j 行的单元格记录了取值为第 i 个列变量值和第 j 个行变量值的实体的频次数。表中央 r 列 c 行的 $r \times c$ 个元素 $n_{ij}(i = 1 \cdots r, j = 1 \cdots c)$ 是表的主体,主体的每一列是自变量取值确定后应变量 Y 的分布,称为条件分布。表的最后一列是每行各单元格频次的和,称为行变量 Y 的边缘和或 Y 的边缘分布。边缘分布是不考虑自变量影响情况下 Y 的分布,

先计算 E_1:当自变量 X 的取值不确定时,为了尽可能准确地预测 Y,减少预测误差,自然应该挑选 Y_j 中的众值作为所有实体的预测值,即选取对应于 Y 的边缘分布的诸 n_{*j} 中最大值 $\max(n_{*j})$ 的 Y_j,来预测每一个实体的 Y 值。按这种方法进行预测,将有 $\max(n_{*j})$ 个实体预测正确,$n - \max(n_{*j})$ 个实体预测错误。因此

$$E_1 = n - \max(n_{*j}) \tag{10-15}$$

再计算 E_2:当自变量 X 的取值已知,例如 $X = X_i$ 时,就要利用表中第 i 列诸 n_{ij} 的分布。为了尽可能准确地预测自变量取值为 X_i 的 n_{i*} 个实体的 Y 值,应挑选第 i 列中对应于诸 n_{ij} 中最大值 $\max(n_{ij})$ 的 Y_j 作为预测值。按这种方法进行预测,对于自变量取值为 X_i 的 n_{i*} 个实体,将有 $\max(n_{ij})$ 个实体预测正确,$n_{i*} - \max(n_{ij})$ 个实体预测错误。至此只考虑

了对应于 X 取值为 X_i 的 n_{i*} 个实体的预测情况，需要考虑全体 X 的 r 个可能取值，因此要对 i 求和，这样

$$E_2 = \sum_{i=1}^{r}(n_{1*} - \max(n_{ij})) = n - \sum_{i=1}^{r}\max(n_{ij}) \tag{10-16}$$

因此

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{\sum_{i=1}^{r}\max(n_{ij}) - \max(n_{*j})}{n - \max(n_{*j})} \tag{10-17}$$

可以证明，公式(10-17)定义的 λ 的值总是在 0 与 1 之间变动。如果 X 与 Y 间无关联，Y 的各条件分布和 Y 的边缘分布应该都是一致的，或者说是成比例的，这样各分布中最大的频次数都处在同一行。因此公式(10-17) 的分子等于零，λ 也就等于 0。而当 X 与 Y 完全关联时，X 完全确定了 Y 的取值，表 10-10 的各行各列中均有一个、且只有一个不为零的频次数，其他单元格中的频次数均为零。因此 $\sum_{i=1}^{r}\max(n_{ij}) = n$，式(10-17) 的分子分母相等，$\lambda$ 就等于 1。

下面我们对表 10-11a 中 136 座墓葬按墓式和墓主人年龄段分布的数据，根据公式(10-17)计算其 λ 值。$\lambda = \frac{(23+19+15)-53}{136-53} = 0.0482$。表明在知道了自变量后，预测误差减少的比例仅 4.8%，说明墓式和墓主人年龄段这两个变量之间的关联是相当弱的。这与 10.3 节根据 χ^2 和 V 进行检验的结论是一致的。

需要指出，λ 值是不对称的，也就是说，如果把 X 与 Y 的关系互换，即 Y 作为自变量、X 作为应变量时，计算得到的 λ 值会发生变化。为了公式表达的对称性，将公式(10-17)重写成

$$\lambda_{xy} = \frac{\sum_{i=1}^{r}\max(n_{ij}) - \max(n_{*j})}{n - \max(n_{*j})} \tag{10-18}$$

如果将 Y 作为自变量，则有

$$\lambda_{yx} = \frac{\sum_{j=1}^{c}\max(n_{ij}) - \max(n_{i*})}{n - \max(n_{i*})} \tag{10-19}$$

λ_{xy} 与 λ_{yx} 是不相等的。对于表 10-11a 的数据，如果以墓式作为自变量，按公式(10-19)计算得到 $\lambda_{yx} = \frac{(23+17+16)-52}{136-52} = 0.0476$。$\lambda_{xy}$ 和 λ_{yx} 虽不相等，但相差也不可能太大，因为它们必竟是同一对变量间关联程度的度量。当自变量和应变量不易分清时，也可以用它们的平均值 $\lambda = \frac{\lambda_{xy} + \lambda_{yx}}{2}$ 作为两个变量间关联程度的度量。在较深入的统计学书中，也介绍怎样用 λ 作为统计量作显著性检验，可近似计算对应于一定 λ 值的显著性水平。某些统计软件也给出 *PRE* 的 λ 值的显著性水平。例如对于表 10-11a 的数据，SPSS 软件给出相应 $\lambda_{yx} = 00476$ 的显著性水平 $\alpha = 0.432$，即应该接受墓式和墓主人年龄段无关的原假设。

10.4.2 *PRE* 的 Goodman and Kruskal's τ 系数

另一个具有 *PRE* 性质的，作为名称变量关联程度的度量是 τ 系数。τ 系数与 λ 系数的不同处在于，在自变量 X 的取值不确定的情况下计算 E_1 时，不再是单一地都用表 10-10 的最右面一列，即 Y 边缘分布的众值来预测全体实体的 Y 值。而是利用 Y 边缘分布中诸 n_{*j} 分布的信息来预测。预测有 n_{*1} 个实体取值 Y_1，有 n_{*2} 个实体取值 Y_2，…… 有 n_{*c} 个实体取值 Y_c。这样对于观测值为 Y_j 的 n_{*j} 实体，预测正确的实体数为 $n_{*j}\times\frac{n_{*j}}{n}$，而预测错误的实体数为 $n_{*j}\times\left(1-\frac{n_{*j}}{n}\right)$。

总的预测错误的实体数是通过对 j 求和得到

$$E_1 = \sum_j n_{*j}\times\left(1-\frac{n_{*j}}{n}\right) = n - \sum_j\frac{n_{*j}^2}{n} \tag{10-20}$$

下面计算 E_2，当自变量 X 的取值已知，为 $X = X_i$ 时，对于 n_{i*} 实体的预测，和上面求 E_1 相似，也考虑第 i 列中诸 n_{ij} 分布的信息，即考虑 $X = X_i$ 时 Y 的条件分布。预测有 n_{i1} 个实体取值 Y_1，有 n_{i2} 个实体取值 Y_2，…… 有 n_{ic} 个实体取值 Y_c。模仿公式(10-20)，对于 $X = X_i$ 的 n_{i*} 个实体，预测错误的个体数为

$$E_2(X = X_i) = n_{i*} - \sum_j\frac{n_{ij}^2}{n_{i*}} \tag{10-21}$$

这样，在自变量 X 的取值已知的条件下，按条件分布预测全体实体的 Y 时，预测错误的总个体数是将公式(10-21) 对 i 求和。

$$E_2 = \sum_i\left(n_{i*} - \sum_j\frac{n_{ij}^2}{n_{i*}}\right) = n - \sum_i\sum_j\frac{n_{ij}^2}{n_{i*}} \tag{10-22}$$

结合公式(10-20)和(10-22)，得

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{\sum_i\sum_j\frac{n_{ij}^2}{n_{i*}} - \sum_j\frac{n_{*j}^2}{n}}{n - \sum_j\frac{n_{*j}^2}{n}} \tag{10-23}$$

可以证明，τ 的取值与 λ 的情况相似，也是在 0 与 1 之间。当 X 与 Y 间无关联时，$\tau = 0$；而当 X 与 Y 完全关联时，$\tau = 1$。另外 τ 也是不对称的，即 τ 也会因 X 与 Y 间自变量的选取不同而有一些差别。

下面仍以表 10-11a 中 136 座墓葬按墓式和墓主人年龄段分布的数据为例计算 τ 值。根据式(10-20)

$$E_1 = 136 - \frac{1}{136}(53^2 + 42^2 + 41^2) = 90.01$$

再根据式(10-22)

$$E_2 = 136 - \frac{1}{45}(23^2 + 12^2 + 10^2) - \frac{1}{52}(19^2 + 17^2 + 16^2) - \frac{1}{39}(11^2 + 13^2 + 15^2) = 88.19$$

因此

$$\tau = \frac{90.01 - 88.19}{90.01} = 0.0202$$

τ 的数值也很小，说明在知道了自变量后，预测误差减少的比例仅为2%，说明墓式和墓主人年龄段这两个变量之间的关联是相当弱的。这与用 V 值和 λ 值度量关联强度作判断时的结论是一致的。τ 的计算过程比 λ 要复杂，但由于它比 λ 更充分地利用了原始数据分布的信息，能更合理地度量两个名称变量间关联的程度。很多统计软件中都包含有计算 λ 与 τ，以及作相应的显著性检验的程序。

10.5　实体按单个名称变量分布的 χ^2 检验

对于实体按数值变量的分布，前几章已有讨论，例如二项式分布、正态分布、t 分布等。名称变量可以取值多个状态，因此也可能遇到分布的问题，例如考虑实体按变量状态的分布是怎样的，是否与某种理论分布一致等。实际分布和理论分布的一致性检验是通过 χ^2 函数来实现的。下面通过实例来说明检验过程。在某地区不同地貌类型的区域调查统计了考古遗址的数量，希望判断遗址的密度和地貌类型间是否有一定的关联，即古代居民在选择居住地时是否偏爱一定的地貌地理环境。

表 10-12 记录了考古调查的结果。

表 10-12　三种地貌类型区域的调查面积和观测到的遗址数统计表

地貌类型	观测到的遗址数 Q_i	调查的面积 S_i (km^2)	调查面积的比例 R_i (%)	H_0 假设下遗址的期望数 E_i
山前平地	26	24.3	32	17.0
沿海地区	9	19	25	13.3
山坡地带	18	32.7	43	22.8
总和	53	76	100	53.1

表 10-12 的第 2、3 列记录实际调查的结果，3 种地貌类型调查区域的面积和观察到的遗址数。这个表和本章前面各节讨论的实体交叉分类频次表是不同的，单元格中记录的面积和面积的比例等不是实体的分类频次。根据所调查的各类地貌区域的面积 S_i，可计算所调查的各类地貌区域面积的百分比 $R_i = \dfrac{S_i}{\sum_i S_i}$（见第 4 列）。第 5 列是在 H_0："古代居民选择居住地时对地貌环境无倾向性"的原假设前提下，计算得到的每种地貌区域的期望遗址数 E_i。计算公式是 $E_i = R_i \times \sum Q_i$。例如山前平地的期望遗址数 E_1 为

$$E_1 = R_1 \times \sum Q_i = 0.32 \times 53 = 17.0$$

同样可以计算 E_2 和 E_3，分别为 13.3 和 22.8。计算得到各地貌类型区域的期望遗址数后，可作假设检验如下：

（1）提出原假设 H_0：古代居民选择居住地时对地貌环境无倾向性。

（2）在 H_0 成立的条件下，计算

$$\chi^2 = \sum_i \frac{(Q_i - E_i)^2}{E_i} = \frac{(26-17)^2}{17} + \frac{(9-13.3)^2}{13.3} + \frac{(26-22.8)^2}{22.8} = 7.11$$

由于所讨论的实例中地貌类型分成 3 类，因此上式计算的 χ^2 服从自由度为 $3-1=2$ 的

χ^2 分布。

(3) 查表 $\chi^2_{0.05}(df = 2) = 5.99 < 7.11$。因此在显著性 $\alpha = 0.05$ 的水平上,拒绝"古代居民选择居住地时对地貌环境无倾向性"的原假设。

χ^2 检验的结论是以稍大于95 % 的置信度判断古代居民选择居住地时对地貌环境有倾向性,但检验的结论一般不能直接推断古代居民更喜欢哪一种地貌环境。对于这个问题的答案需要另外计算比较不同地貌环境区域的遗址密度等。

第十一章　有序变量间的等级相关

前两章分别讨论了数值变量之间的相关关系和名称变量之间的关联关系，处于中间层次的有序变量之间同样可能存在相关问题，有序变量之间的相关又称等级相关。本章将先后讨论表征有序变量之间相关强度的斯皮尔曼(Spearman)相关系数和 Gamma 相关系数，简单介绍 Kendall's τ 系数。本章的最后还将介绍有序变量的百分累加曲线之间的比较。

11.1　斯皮尔曼等级相关系数

为了便于理解，将通过一个实际的例子来说明斯皮尔曼相关系数的定义、计算方法和有关的假设检验。21 世纪初我国的故宫又进行了一次较大规模的修缮工作。修缮的重要内容之一是置换故宫建筑物上大量的琉璃瓦。因为经历长期的冬夏交替和日晒雨淋，不少琉璃瓦胎体上的釉质琉璃有不同程度的剥落。苗建民等(2004)研究了瓦胎上釉质剥落的程度与胎体的孔隙度之间的关系。表 11-1 列出了对 14 片瓦的胎体的气孔率(第 2 列)和釉质剥落程度(第 3 列)的观测数据。虽然表中用百分比作为这两个变量的测量单位，但釉质琉璃层的剥落率是目测的估计值，也可以分成未剥落、极小片剥落一直到严重剥落、完全剥落等级别。气孔率也可以看成有序变量，在 2.2.4 小节中曾提到，高层次的变量总是可以转换为较低层次的变量。表 11-1 的第 4、5 列分别表示这两个变量按各自的大小排序后的序列号。因为总共有 14 个样品，序列号本来应该从 1 开始，每次增加 1，一直到 14。但有时两个或两个以上的样品是等级别的，例如编号为 6 号和 9 号的两个瓦片样品，它们的气孔率排序位置本应定为第 7 和第 8，但它们的气孔率值是相等的，不能分清前后次序，所以它们的气孔率排序都定为第 7.5 位。类似的情况还有编号 2 号和 13 号样品的气孔率排序，37 和 38 号样品的釉剥落率排序等。该表的第 6 列显示两个变量排序次序的差值，称为序差，用 Di 表示。序差可正可负。最后一列是序差的平方项，它总是大于或等于零的正值。

表 11-1　14 片故宫琉璃瓦的胎体气孔率和釉剥落率

瓦片编号	气孔率%	釉剥落率%	气孔率排序 X	釉剥落率排序 Y	序差 D = X − Y	(序差)^2 D^2
38	26	5	1	3.5	−2.5	6.25
37	29	5	2	3.5	−1.5	2.25
11	30	0	3	1.5	1.5	2.25
3	32	25	4	5	−1	1
35	33	0	5	1.5	3.5	12.25
14	37	30	6	6	0	0

续表

瓦片编号	气孔率%	釉剥落率%	气孔率排序	釉剥落率排序	序差	(序差)^2
6	38	65	7.5	10	-2.5	6.25
9	38	80	7.5	12	-4.5	20.25
34	39	80	9	12	-3	9
16	40	40	10	7	3	9
4	42	60	11	9	2	4
2	43	50	12.5	8	4.5	20.25
13	43	90	12.5	14	-1.5	2.25
1	46	80	14	12	2	4

和 = 99

斯皮尔曼相关系数的定义如下：

$$r_s = 1 - \frac{6\sum_i D_i^2}{n(n^2 - 1)} \tag{11-1}$$

式中分式的分子是6倍的序差平方和，n 是样本的容量，对于本例 $n = 14$。

可以证明斯皮尔曼相关系数的取值范围从 +1 到 -1。当两个变量完全正相关时，每个实体的两个变量的等级都相等，因此全部序差均为零，式(11-1)的分式的分子也为零，$r_s = 1$。当两个变量完全负相关时，即实体按两个变量的排序是完全倒序的，第一变量的最低级和第二变量的最高级相对应，通过代数运算可以证明 $r_s = -1$。总之 r_s 的绝对值越接近1，关联程度越高；r_s 的绝对值越接近零，关联程度越低。因此斯皮尔曼相关系数是两个有序变量间等级相关强弱的度量。此外还可以证明，当 $n \geqslant 10$ 时，模仿皮尔逊相关系数的有关公式(9-11)，用 r_s 组成的统计量：

$$t = |r_s|\sqrt{\frac{n-2}{1-r_s^2}} \tag{11-2}$$

同样服从自由度为 $(n-2)$ 的 t 分布。因此可以根据 r_s 和 n 的数值，对有序变量的等级相关作显著性检验。

对于上面故宫琉璃瓦的例子，$\sum_i D_i^2 = 99, r_s = 1 - \frac{6 \times 99}{14(14^2 - 1)} = 0.78$，$t = 0.78\sqrt{\frac{14-2}{1-0.78^2}} = 4.31, df = 14 - 2 = 12$。查 t 分布函数表，得到相应的显著性水平 $\alpha = 0.001$。因此可以以99.9%的置信度判断琉璃瓦胎体的气孔率和釉质剥落的程度是相关的，相关强度为0.78。顺便提到，如果把胎体的气孔率和釉质的剥落程度看成数值变量，计算得到皮尔逊相关系数是0.834，与斯皮尔曼相关系数0.78相差不多。斯皮尔曼相关分析相对于皮尔逊相关分析的优点是，假设检验时不要求应变量 Y 服从正态分布，也不要求各 Y_i 的方差相等等前提条件的成立。对于故宫琉璃瓦的例子，是难以判断釉的剥落率是否服从正态分布的，作等级相关分析更为合适。

对于小样本，当 n 小于10时，公式(11-2)定义的 t 一定程度上偏离 t 分布，这时可查下列的表11-2来确定不同 n 时 t 值所对应的显著性水平和置信度。

表 11-2 小样本情况下等级相关检验中对应于不同的显著性水平和置信度的斯皮尔曼相关系数

显著性水平 α	0.2	0.1	0.05	0.01
置信度(1 - α)	80%	90%	95%	99%
n				
4	0.639	0.907	1.000	
5	0.550	0.734	0.900	1.000
6	0.449	0.638	0.829	0.943
7	0.390	0.570	0.714	0.893
8	0.352	0.516	0.643	0.833
9	0.324	0.477	0.600	0.783

需要指出，使用斯皮尔曼相关系数来表征两个有序变量间的相关强度时要求，同一等级上出现有两个或两个以上实体的次数不应太多，特别是不应出现好几个实体的等级相同的情况。如果样本中存在相同等级的实体，计算斯皮尔曼相关系数的公式(11-1)需要作一些修正。在有的统计学书中给出修正公式，如美国匹茨堡大学周南(R.Drennan)撰写的《Statistics for Archaeologists, A Commonsense Approach》。对于故宫琉璃瓦的例子，作修正后的 $r_s = 0.759$，略小于公式(11-1)给出的未修正值 0.780。

11.2 Gamma 等级相关系数：以陕西史家墓地墓葬分期方案的比较为例

上一节已经提到，当有很多实体处于同一等级时，就不能用斯皮尔曼相关系数来表征有序变量间的相关强度。另外，用斯皮尔曼相关分析方法处理含有大量实体的大样本时，计算工作量很大，在计算机普及前不很方便。因此发展了一种 Gamma 等级相关分析方法。Gamma 等级相关分析把两个有序变量的取值分成数目不多的几个大段，实体同时按照两个有序变量的取值段分类，再统计每类的实体数，并写出类似于第十章的交叉列联频次表。

为了便于理解，还是通过实际的例子来阐明 Gamma 等级相关系数。20 世纪 80 年代我国考古期刊曾先后发表了 6 个对陕西史家仰韶墓地的墓葬进行分期的方案，并曾引起热烈的争论。详细的情况将在本书 17.3 节关于数量方法应用于考古分期研究中介绍。这里仅涉及如何用 Ganma 等级相关分析来定量表述两个分期方案间的相似程度。墓葬分期中的期别属于有序变量，而且一般情况下每一期别含数目相当多的墓葬，因此定量比较两个分期方案的异同程度适宜于使用 Ganma 等级相关分析方法。张忠培(1981)用传统的类型学方法提出了史家墓地的分期方案，后来本书的作者(1985)使用聚类分析的定量方法也建议一个分期方案。有 25 座墓葬在这两个方案都作了分期的。表 11-3 显示了这 25 座墓葬相对于两个分期方案的交叉分类，表中每个单元格决定了每座墓葬的分期位置，第 i 列第 j 行单元格记录了被张定为第 i 期和被陈定为第 j 期的墓葬数目 n_{ij}，n_{ij} 同时也表示单元格的编号。显然 i 与 j 分别表示变量 X 与 Y 的等级。从表中可以看到 $n_{11} = 4$，即有 4 个墓葬被两个方案都定为第一期；$n_{21} = 1$，有 1 个墓葬张定为第二期而陈定为第一

期，等等。

表 11-3 史家墓地 25 座墓葬根据张宗培和陈铁梅两个分期方案的分组表

陈铁梅 Y/张宗培 X	I 期	II 期	III 期	和 (n_{*j})
I 期	$n_{11}=4$	$n_{21}=1$	$n_{31}=2$	7
II 期	$n_{12}=3$	$n_{22}=3$	$n_{32}=2$	8
III 期	$n_{13}=2$	$n_{23}=1$	$n_{33}=3$	6
IV 期	$n_{14}=0$	$n_{24}=1$	$n_{34}=3$	4
和 (n_{i*})	9	6	10	25

为了计算表(11-3)所列数据的 Gamma 相关系数，首先要把表(11-3)中实体两两之间的关系分成三种类型：同序对、逆序对和同分对。设单元格 A 中实体的变量 X 与 Y 的等级是(x_i,y_i)，单元格 B 中实体的变量 X 与 Y 的等级是(x_j,y_j)，现对 3 种关系类型定义如下：

(1) 同序对。

如果对于 X 与 Y 都是 $i<j$，那么 A 格中的每一实体与 B 格中的每一实体均组成同序对。据表 11-3 可见，n_{11} 单元格中的 4 个实体与 n_{22}，n_{23}，n_{24}，n_{32}，n_{33} 和 n_{34} 等 6 单元格中的每个实体都组成同序对。n_{23} 与 n_{34} 的实体也组成同序对，当然还可以组成其他的同序对。同序对的总数用 n_s 表示。

(2) 逆序对。

如果对于 X 有 $i<j$，而对于 Y 有 $i>j$；或者反过来对于 X 有 $i>j$，而对于 Y 有 $i<j$，则 A 格中的实体与 B 格中的实体组成逆序对。例如对于表 11-3，n_{32} 中的两个实体与 n_{13}，n_{14}，n_{23} 和 n_{24} 的实体组成逆序对的例子。逆序对的总数用 n_d 表示。

(3) 同分对。

同分对又分成 3 种情况。

如果对于 X 有 $i=j$，则 A 格与 B 格中的实体组成 X 同分对。同一列各单元格的实体间组成 X 同分对。例如对于表 11-3，n_{11}，n_{12}，n_{13} 和 n_{14} 单元格中的实体相互组成 X 同分对。X 同分对的总数用 T_x 表示。

如果对于 Y 有 $i=j$，则 A 格与 B 格中的实体组成 Y 同分对。同一行各单元格的实体间组成 Y 同分对。对于表 11-3，n_{11}，n_{21} 和 n_{31} 单元格中的实体相互组成 Y 同分对。Y 同分对的总数用 T_y 表示。

处于同一个单元格中的实体，则对于 X 与 Y 均有 $i=j$，它们组成 X，Y 的同分对。X，Y 的同分对只可能存在于同一个单元格的实体之间，而每个单元间的各实体间相互组成 X，Y 的同分对。X，Y 同分对的总数用 T_{xy} 表示。

应该指出同分对的 3 种情况并不是互斥的，X，Y 的同分对是 X 同分对和 Y 同分对的特殊情况。X，Y 的同分对的数目 T_{xy} 已包括在 T_x 与 T_y 之中。因此实体对的总数是 $n_s+n_d+T_x+T_y-T_{xy}$。而且有

$$\frac{n}{2}(n-1)=n_s+n_d+T_x+T_y-T_{xy} \tag{11-3}$$

从实体两两间 3 类关系的定义可以看出，同序对反映两个变量间的正相关，逆序对

反映负相关,而同分对则反映缺乏相关,或变量间的独立性。如果在实体按两个有序变量分布的交叉列联频次表中(例如表 11-3),同序对的数目 n_s 显著超过逆序对的数目 n_d,那么这两个变量间应为正相关;反之,若逆序对的数目显著超过同序对的数目,那么这两个变量间应为负相关。如果 n_s 与 n_d 的大小差不多,那么两个变量间的相关程度很弱。由上面的讨论顺理成章地定义 Gamma 等级相关系数为

$$G = \frac{n_s - n_d}{n_s + n_d} \tag{11-4}$$

由公式(11-4)容易看出,Gamma 的取值范围从 1 到 -1。当列联表中实体间的关系全是同序对时, $n_d = 0$时, $G = 1$,变量间是完全的正相关;当实体间的关系全是逆序对时, $G = -1$,变量间是完全的负相关。当同序对和逆序对的数目相等, $n_s = n_d$ 时, $G = 0$,变量间互相独立,不相关。因此 Gamma 系数是两个有序变量间相关程度的度量。

在讨论了 Gamma 等级相关系数的定义和性质后,回到表 11-3 的例子,分析张忠培和陈铁梅对史家墓地的两个分期方案是否相关,相关的程度有多高。计算表 11-3 的 Gamma 相关系数,完全可以借助于各种计算机软件,但为了演示计算过程,下面进行手工计算。计算的主要内容是确定 n_s 和 n_d。

先计算 n_s,分成几部分计算:

$n_{11}: n_{11} \times (n_{22} + n_{23} + n_{24} + n_{32} + n_{33} + n_{34}) = 4 \times (3 + 1 + 1 + 2 + 3 + 3) = 52$

$n_{21}: n_{21} \times (n_{32} + n_{33} + n_{34}) = 1 \times (2 + 3 + 3) = 8$

$n_{12}: n_{12} \times (n_{23} + n_{24} + n_{33} + n_{34}) = 3 \times (1 + 1 + 3 + 3) = 24$

$n_{22}: n_{22} \times (n_{33} + n_{34}) = 3 \times (3 + 3) = 18$

$n_{13}: n_{13} \times (n_{24} + n_{34}) = 2 \times (1 + 3) = 8$

$n_{23}: n_{23} \times n_{34} = 1 \times 3 = 3$

$n_s = 52 + 8 + 24 + 18 + 8 + 3 = 113$

再计算 n_d,也是分成几部分计算:

$n_{31}: n_{31} \times (n_{22} + n_{23} + n_{24} + n_{12} + n_{13} + n_{14}) = 2 \times (3 + 1 + 1 + 3 + 2 + 0) = 20$

$n_{21}: n_{21} \times (n_{12} + n_{13} + n_{14}) = 1 \times (3 + 2 + 0) = 5$

$n_{32}: n_{32} \times (n_{23} + n_{24} + n_{13} + n_{14}) = 2 \times (1 + 1 + 2 + 0) = 8$

$n_{22}: n_{22} \times (n_{13} + n_{14}) = 3 \times (2 + 0) = 6$

$n_{33}: n_{33} \times (n_{24} + n_{14}) = 3 \times (1 + 0) = 3$

$n_{23}: n_{23} \times n_{14} = 3 \times 0 = 0$

$n_d = (20 + 5 + 8 + 6 + 3 + 0) = 42$

将 n_s 和 n_d 代入公式(11-4),得到

$$G = \frac{113 - 42}{113 + 42} = 0.458$$

Gamma 系数定量地表述了张与陈的两个分期方案间的相关程度。由 G 的数值和符号可以认为,这两个分期方案是正相关的,但相关强度并不大。

顺便指出,如果表 11-3 中的 X 与 Y 都只有两个等级,如下所示:

	X_1	X_2
Y_1	a	b
Y_2	c	d

则 $G = \frac{ad - bc}{ad + bc}$,与第十章的式(10-6)定义的,度量 2×2 四格列联表关联强度的 Yule Q 系数是一致的。因此 Q 系数是 Gamma 系数的一种特殊情况。

11.3 Kendall's τ_b 和 τ_c 等级相关系数

公式(11-4)定义的 Gamma 系数,没有充分考虑同分对对相关系数的影响,因此 Gamma 系数对相关程度的估计偏高。如果将 Gamma 系数公式中的分子 $(n_s - n_d)$ 被所有实体两两成对的总数 $\frac{1}{2}n(n-1)$ 去除,这样定义的相关系数又会对相关程度估计不足。因此 Kendall 提出了两种修正的方法。

(1) Kendall's tau - b 系数定义如下:

$$\tau_b = \frac{(n_s - n_d)}{\sqrt{n_s + n_d + T_x + T_{xy}}\sqrt{n_s + n_d + T_y + T_{xy}}} = \frac{(n_s - n_d)}{\sqrt{\frac{1}{2}n(n-1) - T_y}\sqrt{\frac{1}{2}n(n-1) - T_x}} \tag{11-5}$$

τ_b 相对于式(11-4)的 Gamma 系数而言是在分母中,分别考虑了实体间同分对的影响。对于表 11-3 所示的数据,同分对的数目计算如下:

X 的同分对: $T_x = \sum_i \frac{n_{i*}}{2}(n_{i*} - 1) = 36 + 15 + 45 = 96$

Y 的同分对: $T_y = \sum_i \frac{n_{*j}}{2}(n_{*j} - 1) = 21 + 28 + 15 + 6 = 70$

X, Y 的同分对: $T_{xy} = \sum_i \sum_j \frac{n_{ij}}{2}(n_{ij} - 1) = 6 + 1 + 3 + 3 + 1 + 1 + 3 + 3 = 21$

在分别计算了同序对、异序对和 3 种同分对的数目后,不妨利用计算实体对总数的公式(11-3)来检查各类实体对数目的计算结果是否正确。

$n_s + n_d + T_x + T_y - T_{xy} = 113 + 42 + 96 + 70 - 21 = 300$

$\frac{n}{2}(n-1) = \frac{25}{2}(25-1) = 300$

两种方法计算的总实体对的数目是相符的。肯定计算正确后,将同分对等数据代入式(11-5)得到:

$$\tau_b = \frac{(n_s - n_d)}{\sqrt{n_s + n_d + T_x + T_{xy}}\sqrt{n_s + n_d + T_y + T_{xy}}} = \frac{71}{\sqrt{113 + 42 + 96 + 21}\sqrt{113 + 42 + 70 + 21}} = 0.328$$

（2）Kendall's tau-c 定义如下：

$$\tau_c = \frac{n_s - n_d}{\frac{1}{2}n^2(m-1)/m} \tag{11-6}$$

式中 m 为行数或列数中取小的一个数值。表 11-3 是 3 列 4 行，因此对于表 11-3 有 $m = 3$，计算得到

$$\tau_c = \frac{71 \times 2 \times 3}{25^2(3-1)} = 0.341$$

可以看到，考虑了同分对后的等级相关系数 τ_b 和 τ_c 的数值比 Gamma 系数(0.458)为小。Gamma 系数，τ_b 和 τ_c 不仅定量地度量了两个有序变量间相关的强度，它们也可以作显著性检验，检验方法在较深入的统计学书中有介绍，某些统计软件也能给出检验结果。利用 SPSS 软件对表 11-3 的 Gamma 系数、τ_b 和 τ_c 作检验，所得的显著性水平都是 $\alpha = 0.042$。因此尽管 τ_b 和 τ_c 的数值与 Gamma 系数不一致，只是对相关强度的估计有一定的差别，但检验得到的显著性水平是一致的。检验结果是在 $\alpha = 0.04$ 的水平上认为张与陈的分期方案是相关的，如果将显著性水平定得稍高些，譬如说选定 $\alpha = 0.02$，则将接受“两个分期方案是不相关的”原假设。总之张与陈对史家墓地的分期方案是正相关的，但相关程度并不高。

最后还需要指出，表面上看表 11-3 似乎与第十章名称变量的列联表 10-10 十分相似，每个单元格中记录的都是对应于有关行、列变量值的实体的频次。但这两张表之间是有区别的。第十章名称变量列联表中的行、列变量都是名称变量，使用 χ^2，V，λ 和 τ 等统计量来描述变量间的关联程度。而且对于名称变量列联表，行与行之间，列与列之间是可以互相任意交换位置的。而且交换位置后，χ^2 等关联强度系数的数值不会发生变化。但是对于有序变量的表 11-3，其行(列)的次序反映墓葬分期的次序，行(列)与行(列)间的换位会改变分期方案，因此这种换位是不被允许的。名称变量列联表的假设检验要求表中每个单元的频次数值不能太低，譬如说不小于 5。但是计算 Ganma 等各等级相关系数时，对于列联表中每类实体的频次数并没有限制。

我们还可以进一步分析下面的一张列联表，来说明两种列联表间的差别：

	X_1	X_2	X_3	X_4
Y_1	0	0	n_{31}	0
Y_2	n_{12}	0	0	0
Y_3	0	n_{23}	0	0
Y_4	0	0	0	n_{44}

这张列联表的每行每列都只有一个单元格不为零。如果 X 与 Y 是名称变量，X 完全决定了 Y 的取值，两个变量间是函数关系。其关联系数 λ 和 Goodman-Kruskal's τ 都应该等于 1，X 与 Y 间是强关联。但如果 X 与 Y 是有序变量，凭直观分析就可以看出 X 与 Y 间相关性不可能很高，各种等级相关系数 r_s，*Gamma*，*Kendall's* τ_b，τ_c 等都不可能等于 1。因为这里研究的是两种不同层次的变量间的关系，列联关联强并不表示等级相关也强。列联关联

强表示当自变量 X 已知时可以准确地预测应变量 Y 的取值,等级相关分析的是两个变量变化的方向是否有关系。两种不同层次的变量,研究不同性质的问题,因此表 11-3 与表 10-10 的相似仅仅是表观的。

11.4 两个有序变量百分累加曲线的一致性检验

第三章的图 3-1c 是一张百分累加曲线图,描述性统计了青海乐都柳湾墓地成年女性人骨按年龄段百分比的增长。本节将通过考古实例来讨论两条百分累加曲线的一致性检验。

考古调查了某新石器晚期的墓地。按照墓中随葬品的多寡和质量分成 76 座富人墓和 136 座穷人墓,并统计了墓主人的死亡年龄。表 11-4 列出墓葬按贫富情况和死亡年龄段分组的数据。

表 11-4 某墓地墓葬按照贫富情况和死亡年龄段的分组统计

年龄段	人数	人数	百分数	百分数	累计百分数	累计百分数	累计百分数之差
	富	贫	富	贫	富	贫	
婴儿	6	23	0.079	0.169	0.079	0.169	0.090
儿童	8	21	0.105	0.154	0.184	0.323	0.139
少年	11	25	0.145	0.184	0.329	0.507	**0.178**
青年	29	36	0.382	0.265	0.711	0.772	0.061
壮年	19	27	0.250	0.199	0.961	0.972	0.010
老年	3	4	0.039	0.029	1.000	1.000	0.000
总和	76	136	1.000	1.000			

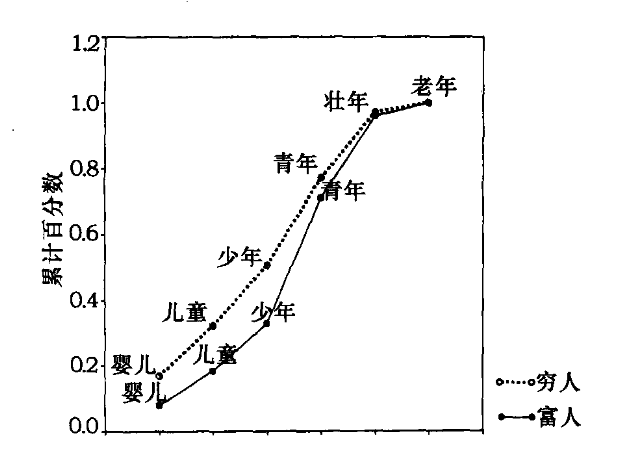


图 11-1 某墓地墓葬按照贫富情况分类的两条死亡年龄段百分累加曲线

图 11-1 是分别反映贫富两类墓葬按墓主人死亡年龄段分布的两条百分累加曲线。由图可见在低年龄段穷人的死亡率高于富人的,穷人墓葬按死亡年龄段分布的百分累加曲线高于富人墓葬的。希望判断,所观察到的贫富间死亡率的差异属于随机的涨落,还

是有统计意义的，即我们要检验两条死亡年龄段百分累加曲线的差别是否显著，对墓地所属的氏族这个总体而言，人的寿命和他们的财产状态之间有没有关系。

柯尔莫高罗夫和斯米尔诺夫提出了一种检验两条百分累加曲线一致性的方法，把两条百分累加曲线的最大差值作为一个判别量。在我们的例子中最大差值是0.178（见表11-4，用黑体字表示），对应于“少年”段。

柯尔莫高罗夫和斯米尔诺夫的检验标准是将最大差值与 χ_α 相比，χ_α 按下式计算：

$$\chi_\alpha = K_\alpha \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \tag{11-7}$$

式中 K_α 是一个常数，随显著性水平 α 而变化。$K_{0.05} = 1.36$，$K_{0.01} = 1.63$ 和 $K_{0.001} = 1.95$。

n_1 和 n_2 是组成两条百分累加曲线的样本的容量，在本例中是76和136。检验过程如下：

(1) 提出原假设 H_0：贫、富两个总体的各年龄段死亡率百分累加曲线间无显著差别

(2) 取显著性水平 $\alpha = 0.05$，计算得到

$$\chi_{0.05} = 1.36\sqrt{\frac{76 + 136}{76 \times 136}} = 0.196$$

(3) $\chi_{0.05} = 0.196 >$ 最大差值 $= 0.178$，因此在 $\alpha = 0.05$ 的显著性水平上接受原假设 H_0。

检验结论是：在 $\alpha = 0.05$ 的水平上没有观察到该新石器晚期墓地所属氏族人口的寿命和人员的财产情况之间有明显的关系。

前面我们仅介绍了柯尔莫高罗夫和斯米尔诺夫的检验过程，并没有讨论为什么可以这样处理。这里仅指出，在实际应用中需要注意两个前提：(1) 它仅适用于有序变量（数值变量可以转换成有序变量）；(2) 样本的容量不能太少，n_1 和 n_2 都应该大于40。

第十二章　抽样问题和考古样本的采集和评估

12.1　抽样问题在总体参数估计中的重要性

在讨论抽样问题前,先简要回顾第五章以来讨论统计推断中关于总体和样本的关系问题。在那些章节讨论了怎样用样本的方差和平均值来估计总体的方差和平均值,用样本的比例数去估计总体的比例数,用样本的相关系数去估计总体的相关系数等,还讨论了估计的置信度和精确度。在所有这些讨论中自始至终默认了一条原则,即所采集或观测得到的样本能"代表"总体,样本应该从总体中随机抽取的,要求样本与总体具有相同的分布。如果总体的参数是已知的,可以用假设检验的方法来检验某个样本是否从总体中随机产生的。一般情况下总体是未知的,正是我们的研究对象,而且是需要根据由观测资料所组成的样本来研究。因此样本的采集和抽取都必须服从一定的准则,使用科学的抽样方法,以保证样本具有良好的代表性。在考古学的研究中,考古学家所掌握的资料,往往是通过考古学特殊的具有自身学科特点的方法所获得的。当利用局部的考古资料对总体、对更大地域范围、更大时间跨度的古代社会作推断时,也必须考虑我们所掌握的考古资料,考古样本能否"代表"总体,代表古代社会。如果缺乏"代表性",那么在对总体的推断中尽管正确使用了第 5-11 章介绍的统计学的各种方法,也难免会得出不完全正确、甚至错误的结论。

下面通过一些实例来说明抽样问题的重要性。先看一个也许是有点极端的假想例子。某个社会统计机构派出两位调查统计员调查统计北京市就业人员的平均工资。其中的一位到居民小区和建筑工地调查了几百位物业管理人员、施工工人、保安员和电梯驾驶员,得出月平均工资为 600 ± 180 元,另一位专门找了几百位写字楼中的白领,他们的平均工资为 4000 ± 1500 元。显然这两个样本对于"北京市就业人员的平均工资"这个总体都缺乏代表性,两个样本中个体的工资分布是不一致的,与总体中的工资分布也不一致。因此这两个样本都不是无偏的样本。为了正确统计北京市就业人员的平均工资,需要制定一个经仔细考虑,并符合抽样基本原则的调查方案。作为不科学抽样导致错误结论的例子,还可以举出美国《文学文摘》(*Literary Digest*)杂志组织的民意调查错误预测了 1936 年的总统选举,很多统计学书中都提到这个例子。《文学文摘》曾于 1932 年组织了民意调查并正确地预测了当年的总统选举,预测的选举票数和实际的票数相差小于 1%。1936 年《文学文摘》根据所掌握的电话黄页和一些俱乐部的名册发出了 1000 万封调查信,收到 200 万份回复。回复的信中压倒多数支持共和党候选人兰登,但实际的选举结果是民主党的罗斯福得到了 61%的选票,而兰登才得到 39%的支持。为什么《文学文摘》的预测偏离真实如此远呢?因为当时拥有电话或参加各种俱乐部的是富人的比例大,而且对调查作出回应的更是对美国 30 年代经济衰退时期罗斯福的政策不满的那部分富人。

因此这200万份回复所组成的样本是有明显的倾向性的，没有按应有的比例反映中等收入者和穷人的意见，样本离“无偏”甚远。1936年，美国盖洛普民意学会却正确地预测了罗斯福的胜利。因为盖洛普根据选民的地区、性别、年龄，特别是收入和财富情况按相应的比例发出了调查信，盖洛普的调查接近于本章后面要介绍的分层抽样。

总之，科学地抽样对统计推断是十分重要的，抽样方法本身也属于统计学中的重要组成部分。

12.2　抽样方法简介

为了样本具有良好的代表性，研究发展了多种科学抽样的方法。下面将介绍简单随机抽样，分层抽样，集团抽样和系统抽样等常用的方法。实际工作中采用哪种抽样方法以及所抽取的样本应包含多少个实体。取决于总体的性质，研究的目的，所要求的判断置信度以及研究者可能掌握的研究时间，人力和经费等，需要统筹考虑这些因素。这些抽样方法的基本原则也是适合于考古学研究的，可以利用这些基本原则来考察实际考古资料的代表性。至于具体的抽样方法与考古研究关系较密切的是系统抽样方法，将在12.2.4节中讨论。

12.2.1　简单随机抽样

简单随机抽样是最基本的抽样方法。我们先介绍随机数和随机数表的概念，设想有一个口袋，其中放入刻有0，1……9的球。袋中每个号码的球的数量是相等的，譬如说都是20个，那么袋中共有200个球。现从中任意抽取一个，记录球的号码数后，将球放回，混匀。然后再抽一个，按次记录，再放回。如此一直进行下去，就得到一张随机数表，如表12-1所示。表12-1包含了从0到9共1152个随机数，是1152次抽取结果的记录。上述建立随机数表的过程实际上就是回放的简单随机抽样过程。表12-1所含的随机数数量比较小，很多统计软件，如SPSS等都可以产生随机数。严格地说统计软件产生的随机数称为伪随机数，但不影响我们的使用。表12-1所列的随机数和一般说的随机数都是指均匀分布的随机数，某些统计软件还可产生其他分布的随机数，例如按标准型正态分布的随机数等。

表12-1　0—9的1152个随机数表

行\列	1	2	3	4	5	6	7	8
1	9895	9659	8996	0938	5774	8057	0644	0152
2	0262	9271	0058	7705	0499	7138	1694	3730
3	2456	0629	7789	6914	5739	2070	2838	2552
4	4110	8905	9003	7969	6713	9146	8760	1189
5	1170	2789	8101	9133	8613	2652	7050	1187
6	5218	7527	2898	8788	6991	4744	1048	1130
7	8129	6859	5443	6211	0826	0953	1485	7849
8	9482	3617	8154	7629	6036	3808	9799	4215
9	3807	1837	5403	6543	1913	4482	8862	2105

续表

行\列	1	2	3	4	5	6	7	8
10	3394	4006	4642	3112	0848	3433	5376	6754
11	9973	6613	2782	3003	5167	3397	7029	6075
12	0629	6396	8754	6679	0311	9130	2688	1025
13	1979	9928	4464	0175	5316	6178	1458	6863
14	5810	4788	3510	9107	4945	4720	7031	6181
15	9106	7178	6186	4216	1037	9040	5091	9767
16	8720	5198	2417	0081	6979	4115	9921	7131
17	9321	8550	4375	8826	3496	5735	5763	6335
18	1320	7097	8529	2908	9390	2483	4785	0278
19	3578	3175	8943	2230	8147	9158	9953	0544
20	5794	2418	1574	9371	7657	1844	6904	4788
21	0613	7837	5338	6056	9835	5272	7501	8586
22	4486	6922	3026	6875	4655	0325	0890	0298
23	8031	6031	4584	6007	5015	6965	3182	2171
24	8296	8604	1880	7050	9835	6794	2210	7759
25	6846	1692	3979	2019	2514	9075	1500	5805
26	1974	5609	1505	8869	9909	6199	1267	3680
27	2912	1389	9941	0395	8868	8099	2638	9219
28	8096	4186	7808	9588	9931	9218	4368	7952
29	0408	0484	3211	1370	4163	4764	7958	9927
30	9757	9006	9469	9324	3464	4539	5434	3477
31	1887	2470	7381	4843	1542	9309	0800	0405
32	0656	9560	9287	5777	3021	4969	9316	8470
33	4471	6851	9722	5735	8011	5551	3035	9387
34	4489	0641	6784	3715	2703	8509	2459	7988
35	2700	5940	5153	7685	4689	7786	1583	7625
36	4248	9670	6768	4740	5733	4504	7859	5828

利用随机数表的帮助,可以进行简单随机抽样。例如对图 12-1 所示的一块地,我们希望在上面随机分布地打 20 个探孔。怎样随机地选择探孔的位置呢。第一步把这块地按照长和宽的比例分成 14 × 7 = 98 个面积相等的接近正方形的小地块。分割地块的行数和列数可以变动,但小地块的总数至少应是探孔数的若干倍。这里按 14 × 7 分割是为了使小地块的总数小于 100,这样取两位随机数就可以确定探孔的位置。对这些正方形地块先按行、再按列编号,从第 00 号编到第 97 号(图 12-1 中只显示了被选中的小地块的编号)。第二步从表 12-1 的任一行任一列开始,譬如从第 7 行第 6 列开始,每 2 个 2 个的顺序取数。结果为 09,53,14,85,78,49,94,82,……等共 20 个数,相应编号的正方形地块被选中,并在图 12-1 中标出。在顺序取数时,如果遇到大于 97 的两位数(本例中曾遇到“99”),或某一个两位数重复出现时,应予舍弃,并继续顺序取数。取数的顺序可以先按行,也可以先按列,也可以每跳一格取数。这样就随机确定了 20 个探孔的位置。

								08	09				
14	15		17										
	29							36		38			
42							49				53	54	
				60									
						76		78			81	82	
	85									94			97

图 12-1　利用随机数表从 **98** 块面积相等的地块中随机选取 **20** 个地块

简单随机抽样方法在社会现象研究，在工厂的产品检验中经常被应用，但在考古调查的探孔和探方布局中应用受限。因为简单随机抽样安排探孔点位置的结果，经常会出现有的区域探孔点密集，而在另外的区域缺少探孔点的情况。这在图 12-1 中也有表现，右下区和左面偏上区被选中的小地块的密度大，而右上和左下区域的地块很少有被选中的。关于考古调查中探孔的布局经常是使用系统抽样方法，这将在后面讨论。

12.2.2　简单随机抽样中样本容量的确定

在上面的例子中，并未说明为什么要打 20 个探孔，可能是因探孔目的以及经费时间等因素而定的。如果抽样的目的是为了根据样本对总体的参数作估计，那么对样本的容量就有确定的要求了。下面分别讨论估计总体平均值和比例数时怎样根据估计中的置信度 $(1-\alpha)$ 要求和估计中所能容忍的偏差 d 来确定样本的容量。

（一）对总体平均值的估计。

在 5.3.3 小节曾给出对总体平均值置信度为 $(1-\alpha)$ 的区间估计的半宽度为

$$d = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{12-1}$$

式中 $Z_{\frac{\alpha}{2}}$ 是置信度为 $(1-\alpha)$ 时的置信阈，一般 α 取 0.05，对标准型正态分布 $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$。$\sigma$ 是总体的标准差。如果 σ 未知，可以先抽取一个小样本，测量计算小样本的标准差 s，用 s 替代 σ。式(12-1) 中的 n 是需要确定的样本的最小容量。式(12-1)可以改写为

$$n = \frac{Z_{\frac{\alpha}{2}}^2 \times \sigma^2}{d^2}，或者\ n = \frac{(1.96)^2\sigma^2}{d^2} = \frac{3.84 \times s^2}{d^2} \tag{12-2}$$

式中的 d 是估计值和真实值之间可以容忍的差别。

实例　要求以 0.95 的置信度估计中国男子的平均身高，估计值和真实值之间可以容忍的差别不大于 0.2 cm，计算最少需要抽取多少个个体 n。

为了利用公式(12-1)求 n，首先要知道中国男子身高的标准差 σ。如果 σ 未知，则先随机抽取一个小样本，例如先抽取 50 人，测定其标准差，譬如说得到 s = 5cm。代入式(12-2)得

$$n = \frac{3.84 \times 5^2}{0.2^2} = 2400(人)$$

随机抽取和测量了 2400 个男子的身高后，可以再计算这个大样本的标准差 s_0，验证与 5cm 有多大偏离，是否可以容忍，考虑是否需要再增加抽样数量。

中国男子的身高这个总体包含了好几亿个个体，接近于无限总体。但有的情况下，

总体包含的实体数 N 不很大。这种情况称为有限总体,考古总体多数为有限总体。对于有限总体公式(12-2)要作相应的修正。

$$n = \frac{N \times Z_{\frac{\alpha}{2}}^2 \times \sigma^2}{d^2(N-1) + Z_{\frac{\alpha}{2}}^2 \times \sigma^2} \tag{12-3}$$

当总体包含的个体数 N 很大时,分母上的第二项可以忽略不计,而且 $N \approx (N-1)$,公式(12-3)还原成公式(12-2)。

(二) 对总体比例数的估计。

第八章的公式(8-2)给出

$$\text{总体比例数的估计误差} = d = Z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})/n} \tag{12-4}$$

这个式子可改写成

$$n = Z_{\frac{\alpha}{2}}^2 \frac{\hat{p}(1-\hat{p})}{d^2} \tag{12-5}$$

式中的 $Z_{\frac{\alpha}{2}}$ 依旧是按正态分布确定的置信阈,$\hat{p}$是样本的比例数。利用式(12-5)可以在选定的置信度$(1-\alpha)$和容忍度 d 的条件下,计算为估计总体的比例数 p 最少所需抽取的实体数目。如果样本的比例数 $\hat{p}$未知,可以先随机抽取少量个体,用小样本的 $\hat{p}$值代入公式(12-5)进行计算,待正式抽样后再验证小样本的比例数是否可以接受。

对于有限总体,当抽样的个体数 n 与总体所包含的个体数 N 可比时$\left(\text{例如}: \frac{n}{N} > 0.05\right)$,公式(12-5)也要作修正:

$$n = \frac{N \times Z_{\frac{\alpha}{2}}^2 \times \hat{p} \times (1-\hat{p})}{d^2(N-1) + Z_{\frac{\alpha}{2}}^2 \times \hat{p} \times (1-\hat{p})} \tag{12-6}$$

实例:这里我们转引谢衷洁(2004)曾讨论的关于 2003 年爆发的非典型性肺炎死亡率的例子。据我国权威人士统计,非典患者的死亡率不超过 6%。但香港报刊报道死亡率为 10% ~ 12%,而世界卫生组织的专家估计死亡率达 15%。如果要求估计死亡率的置信度为 95%,估计误差不超过 1.5%,至少需要由多少名非典病人组成的样本,才能达到这个要求。利用公式(12-5)计算样本容量 n 时,应该知道 $\hat{p}$的数值。可是 3 个来源对死亡率的估计有明显差别,我们取 $\hat{p}$ = 0.15%,因为这样计算得到的 n 值最大,比较保险。代入式(12-5)计算

$$n = \frac{(1.96)^2 \times 0.15 \times 0.85}{(0.015)^2} = 2177(\text{人})$$

即至少需要抽取一个含有 2177 名非典患者的样本。我们知道我国的非典患者为 5327 人,加上加拿大,新加坡等地的患者,完全有足够的非典病例个案对死亡率作出 95%的置信度,误差小于 1.5%的估计。但上面 3 个资料分析来源对死亡率的估计的差别远大于 1.5%。出现这种情况可能是因为 3 个来源使用的样本是不一样的,病人所受到的治疗方案不一样。从统计学的角度分析,3 个来源所使用的样本中包含了对总体而言不是无偏的样本。

12.2.3 分层抽样和集团抽样

分层抽样又称分类抽样。在本章的引言中提到了统计北京市就业人员平均工资的

例子。无论是调查居民区和建筑工地物业管理人员、施工工人、保安员和电梯驾驶员的工资，还是调查写字楼中白领工资的数据，它们对于总体而言都是有偏的样本，缺乏代表性。因为总体包含了各层次工资水平的人群。对于这种本身存在层次的总体，很多情况下估计其某个数值型随机变量的平均值是没有意义的，例如估计幼儿园全部人员的平均身高或平均年龄等是毫无意义的。但有时仍会被要求对这种分层总体的某个变量平均值作估计，总体的平均值仍有一定意义。例如对北京市全部就业人员平均工资的估计应该是预测北京市市场的一个参考因素。对于这类总体平均值的估计，需要用分层抽样的方法来抽取样本，先把就业人员分成从老总、高层白领、一般职员、……到清洁工等各种层次的实体，再从各层次中抽样。

分层抽样的基本思想和程序是，首先根据我们对总体已有的知识，将总体的 N 个个体分成若干组，每一组 N_i 个个体之间的差别应该尽量小，用少量抽样得到各组方差的粗略值 s_i^2，然后以 $\frac{N_i}{N}$ 为权计算总体的加权方差。第二步按照估计置信度 $(1-a)$ 和容忍度 d 的要求和加权方差值计算总共需要抽取多少个个体 n。再按照各组标准差的粗略值 s_i 和权 $\frac{N_i}{N}$ 计算每组需要抽取的个体数的比例数，最后将总共需要抽取的个体数 n 按比例分到各组。分层抽样适用于组内差别小而组间差别大的总体。

在考古学研究中分层总体不常见，而且即使遇到了分层总体，也难以估计各组的权重。分层抽样在考古研究中难得应用，也许把分层的各组作为若干个独立的总体来对待更为合适。因此我们对分层抽样的讨论仅限于上述的基本原理。

与分层抽样对立的是集团抽样，后者适用于组内差别大而组间差别小的可分组总体。还是以北京市就业人员平均工资的统计为例，也许选几个集团公司的全体人员作为调查对象会得到比较接近真实的结果。因为每个集团中都有老总、白领、蓝领、一般职员，以及保安、清洁工等各类人员，而且各类人员的组成比例也应比较接近全市各类就业人员的组成比例。集团抽样的优点是省时省钱，得到结果快。但需要注意集团本身的代表性，如果选取一个行将破产的集团作调查，调查结果将显著偏离总体的平均状态。集团抽样在考古研究中的应用似乎也不普遍。

12.2.4　系统抽样和考古调查中的探孔布局和探方尺寸问题

系统抽样的程序是这样的。先把总体的 N 个个体按某个因素排列。确定总共计划抽取的个体数 n，这样抽样间隔 $r=\frac{N}{n}$，在第一间隔的 r 个个体中任选一个为抽样的起始点，向后面每隔 r 个个体抽样。总共将抽取 n 个个体组成样本。由于在第一间隔中抽样起始点的选取是任意的，总体中每个个体就有同等的概率被抽取，因此系统抽样仍不失其随机性。决定总体个体排列的因素和抽样研究的目的可以有关，也可以无关，总体也可以按随机数排列。需要注意的是，如果排列后的个体的某种属性的取值有周期性，而抽样研究的目的又与该属性有关联时，系统抽样的样本可能会有系统偏差。例如调查某公园全年每日游客的平均数，按 7 天间隔抽样。有可能都是抽取每逢周日的游客数，样本的平均值会显著偏高，另一种情况是抽样日都不是周末，例如每个星期三，样本的平均值

可能会稍偏低。因此在这个调查中系统抽样的间隔不应该取 7 天，系统抽样的间隔周期与被研究对象的变化周期不应该同步。

系统抽样方法在考古调查的探孔布局方案中得到应用。例如要在一定面积的地块上钻一定数量的探孔，系统布置探孔比随机布置有更高的概率发现遗址。Champion (1996) 曾对此作了专门的研究，他计算表明，在探孔总数确定的条件下，按正三角形布置探孔比按正方形布置，探孔处于遗址部位的概率更高（这里 Champion 把墓葬等有一定面积的遗存也称为遗址）。Champion 在一块 357 × 1429 米的地块上按正三角形布置了探孔（见图 12-2）。

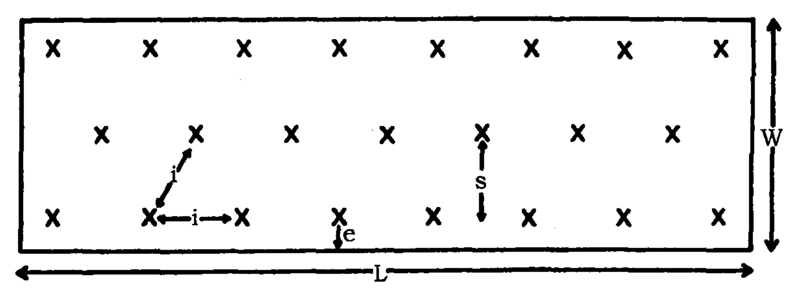


图 12-2 探孔按正三角形布局的一个例子（示意图）

图中地块的宽度 $W = 357m$，地块长度 $L = 1429m$，正三角形的边长 $i = 88.3m$，探孔的行距 $s = i\frac{\sqrt{3}}{2} = 76.5m$，探孔的行数 $t = 5$，探孔至长边的距离 $e = \frac{1}{2}(357 - s(t-1)) = 25.5m$，即 $e = \frac{s}{3}$。在这个调查中他发现了一个遗址，它的最大线性长度是 128 米。另外 Champion 计算表明，在上述的探孔布局方案中，只要遗址的中心处于上面的地块中，并且其直径大于 102 米，那么至少会有一个探孔位于该遗址的位置上。但是探孔处在遗址的位置上，并不能保证一定能发现遗址，特别是当遗物在遗址的分布是稀疏的情况。即使探孔打在遗址的部位，但因探孔本身的面积不大，而遗物密度又小，探孔取样中完全有可能见不到任何一件遗物，从而不能识别探到了遗址。“探孔处于遗址部位”和“探孔确定了遗址的存在”是两个不同的概念。真正发现遗址的概率还与探孔（探方）本身的面积，以及遗物在遗址范围中的分布密度和分布模式有关。Champion 对这块地块进行了全面的考古发掘，然后他根据实际发掘的资料，反推在上述的探方布局条件下探方本身的面积是怎样影响探测到遗址的存在。其研究结果用图 12-3 中偏下面的一条曲线来显示。

该图的 Y 轴显示发现遗址的概率，X 轴表示探方的面积。可以看到随探方面积的增大，遗址被发现的概率也增加。这应该是不言而喻的，但值得注意的是图上曲线的增长有一个转折点，当探方的面积扩大到一定程度（10—15 平方米）后，遗址被发现的概率已接近 90%，再增大探方的面积，遗址被发现的概率增加就很慢了。也就是说在地块的考古调查中探方不必开得太大，以节省工时。考古调查中探方的密度和面积也因调查的目的而异，为寻找农业时代的聚落或石器时代游牧的营地，后一种情况下探方的密度应安排高些、探方的面积也应大些。总之统计学中系统抽样的某些思想可为考古调查中布局探孔和探方时参考，既要保证不出现重要的遗存如墓葬等被漏查的尴尬的情况，又要追

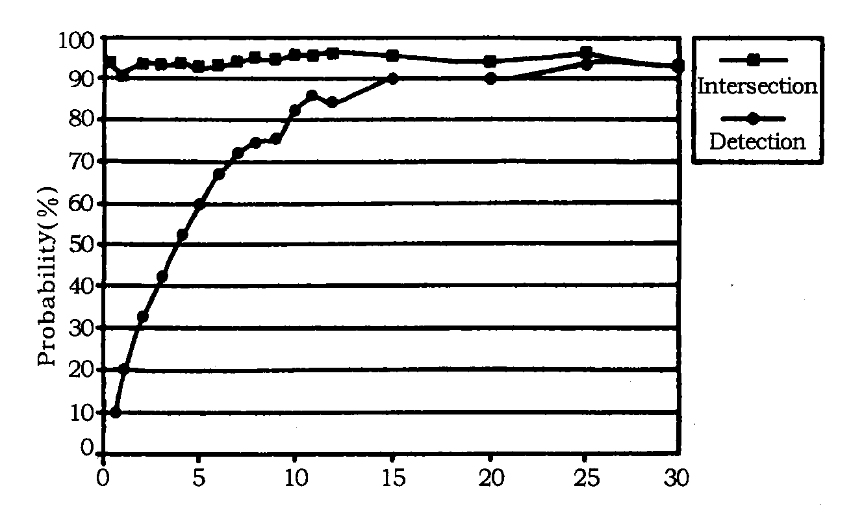


图 12-3　探方的面积和探方确定遗址存在概率的关系曲线

求调查的效率和降低费用。

12.3　考古研究中样本与总体关系的某些特殊问题

阅读本书第五章到第十一章的内容,读者会明确地感觉到,样本与总体间的关系问题,是各类统计推断的中心问题。统计推断遵循着这样一个逻辑:样本来自一个比它容量大的多的,甚至是无限容量的总体,样本应该是从总体中随机抽样得到的,是总体的一部分并能"代表"总体。样本是我们所能掌握的数据资料,而我们真正感兴趣的是总体。统计推断给出了一系列的方法,技术帮助我们根据样本来推断总体的性质,并且给出了推断的置信度和误差。考古学是利用考古发掘所得的古代遗存来推断古代人类社会情况的科学,当我们使用统计推断的各种技术,从遗物遗存来推断古代社会时,我们也许会问,我们所掌握的实物资料是从怎样的总体中来的,它们是不是无偏的随机样本,能否代表总体。这类问题的答案,有时是直接明晰的,有时却不是那么清晰。下面通过三个具体例子进一步探讨考古样本和考古总体的关系问题。

1. 考古学家发掘了一个石器时代遗址,发现有几万件甚至十几万件石器石片遗存,随机采集了几千件。这里样本与总体的关系是明确的,总体是遗址中全部石器石片,而抽取的几千件石器石片就是样本。通过对这几千件石器石片的分类,石料质地的分析和几何尺寸的测量,统计学的方法可以帮助正确推断该遗址全部石器石片中各类石器的百分比,使用各类石料的百分比,各类石器的平均尺寸等,而且能为这些推论赋以定量的置信度和误差估计。

2. 第二个例子是对某个青铜时代聚落遗址进行了考古发掘,确定它分为前后两期。从两期的堆积中都发现了大型陶罐(以下简称陶罐),推测为储藏粮食使用。经对部分陶罐的修复和测量,得知前后两期大型陶罐平均容积 90% 置信度的估计区间分别为 35 ± 5 和 52 ± 4 升。当然我们可以进一步做 t 检验,判断两期陶罐总体的平均容积有没有显著差异。问题是应该怎样理解两期陶罐的总体。一种可能是将该聚落遗址前后两期曾经

使用的全部大型陶罐设想为两个总体。当然不可能复原全部破损的陶罐,这仅是两个假想的总体,而且是实体数有限的总体。但是根据所复原和测量了的两期部分陶罐的容积(它们是从总体中抽取的样本),t检验可以以相当高的置信度推断,认为该聚落两期陶罐的容积发生了明显的变化。考古学家也许更感兴趣的是,该聚落所在地区青铜时代的陶罐前后期是否发生了变化。这样就把该地区同类青铜文化各聚落前后两期的全部陶罐设想为两个总体。那么前面局限于对某个聚落所作统计推断所得到的结论能否推广到地区呢。这就要考察对于陶罐的容积而言,该聚落遗址是否有代表性,典型性。这类似于集团抽样的情况,需要分析集团抽样的样本是不是无偏的。为此需要在该地区另找若干处青铜聚落遗址,分析比较两期陶罐容积在遗址间的一致性。如果一致性检验被通过,各遗址陶罐容积数据就可以合并处理,并可以对该地区青铜时代全部聚落前后两期陶罐的容积变化作统计推断。如果一致性检验被否定,说明各聚落的陶罐可能不属于同一类型,谈论各聚落全部陶罐也就失去意义了。

顺便指出这里的推断仅限于两期陶罐的容积。考古学家也许会推测这类大型陶罐的功能是为存放粮食,并依据后期陶罐容积的增大进一步推测是反映家庭人口的增长。这种推测也许是对的。但是统计推断两期陶罐的容积变化的置信度与推测两期家庭人口增长的可靠性的关系并不是直接的。

3. 在某个地区进行了全面的考古调查和发掘,发现两期聚落遗址各10个,并测量了面积。因为考古调查是全面详尽的,极少有聚落被遗漏。这种情况下样本和总体是接近一致的,再讨论“根据样本来推断总体”似乎显得勉强。那么前面介绍的各种统计推断技术还能否用于这样的数据呢。技术层面的答案是肯定的,因为照样可以计算两期聚落的平均面积和标准差,譬如得到$0.6 \pm 0.4km^2$和$1.1 \pm 0.4km^2$。也可以作t检验,有91%的置信度判断两期聚落的面积有差异。还可以在不同的置信度下计算两期聚落的面积差值的估计区间,如果取置信度为80%,那么差值的估计区间是$0.5 \pm 0.36km^2$。但是这样处理数据、作统计推断还有没有实际的考古意义呢?显然这种情况下的总体完全是假想的,从总体和样本关系的角度看,所进行统计推断的目的似乎是模糊的。但是我们认为,数据的统计处理还是有意义的,其实际意义在于说明一个事实,即这个实例中数据的容量是足够大的。因为数据容量大,我们能以相当高的置信度来判断两期聚落面积的差别,或者说所观测到的差别不太可能是因为我们观测的聚落数太少和由此产生的随机涨落所引起的。这给考古学家对所掌握数据资料的质量以信心,甚至进一步帮助他们探讨聚落平均面积的增大与经济的发展、人口的增加的关系等。另一种情况是,如果两期聚落遗址的平均面积和标准差不变,但两期的遗址数不再是各10个,而仅为各5个,这样t检验的结果会认为两期的面积未见显著差别。说明所研究的聚落数目太少了,两期聚落平均面积的比较中存在较大的不确定性,当然更没有意义去讨论聚落面积与经济,人口的关系了。总之,即使对总体进行了较为全面的考古调查和发掘的情况,我们依然可以使用统计推断的技术和置信度等概念,并有助于判断所掌握数据资料的数量能否作为高置信度讨论的依据。

第十三章　SPSS 统计软件包应用简介

统计学作为随机数据处理,分析和推断的一门学科往往涉及大量的计算,特别是多元统计分析,因此统计学的普及应用是与计算机的发展分不开的。在计算机发展的早期,就有软件公司和学术单位编写了各种通用的和专用的统计软件,其价格从几十到几千美元不等。著名的通用的统计软件有 SAS,SPSS, MiniTab,Statistica 等,另外专用于医学和生物学的有著名的 BMDP 软件,专用于聚类分析的 CLUSTAN 等。鉴于统计学在考古研究中的广泛应用,已编写出版了各种专用于考古研究的统计软件包,有的已成为商品。例如英国伦敦考古研究所编写的 The Institut of Archaeology Data Analysis Package, 美国亚里松纳大学人类学系的 The Archaeologist's Analytical Toolkit, 英国兰彻斯特大学编写的 ASP 和澳大利亚悉尼大学人类学系编写的用于多变量考古分析的 MV-ARCH。此外有专用于考古资料空间分布研究的 ARCOSPACE(丹麦 Åarhus 大学编),专用于考古单元排序用的 Numerical ordination and seriation package (法国 ROZOY 公司)。这些专用的考古分析软件一般均比较小,它们的价格在 100 美元以下,但在我国不容易得到。

在我国得到普遍应用的统计软件是 SAS,SPSS 和 MATLAB 中的 STATS 部分。SPSS 的全名是 Statistical Package for Social Science,中文称为社会科学统计软件包。它是目前国际上得到最广泛使用的统计分析软件之一。SPSS 最早是为大型计算机开发的,后来出现了个人用的微机 DOS 版本,但 DOS 版本要求用户学习记忆各种命令,过程和一些语法规则,自己编写简短的程序。90 年代初出现了 SPSS 的 Windows 版本,通过菜单,对话框和图标按钮来完成操作,使用非常方便。SPSS 的功能很多很强,与其他软件间有方便的数据转换和传输接口。

SPSS 虽然功能强大,但它的微机版本所占磁盘空间并不多,仅 100 多兆字节。SPSS 软件是不断更新版本的,非最新版本的 SPSS 程序往往可以免费下载或允许有限时段的使用。

国内出版的专门介绍 SPSS 软件的功能和使用方法的专著、教材和手册等不下几十本。但一般都是几十万字的篇幅,读者阅读费时较多。本章将用很短的,约一万多字的篇幅介绍 SPSS 软件的基本使用方法以及与本书内容有关的功能,接近于使用手册的性质,希望帮助读者在很短的时间内就能入门使用。

本章所介绍的内容是根据 SPSS11.0 的版本编写的。

13.1　数据文件的建立、编辑和数据的预处理

(一) SPSS 数据文件的建立。

使用 SPSS 软件第一步是建立数据文件。其格式与一般的 $r \times c$ 交叉列联表相似。但它规定每一行代表一个实体或个案(case),而每一列的数据对应于某个变量对所有实体

的取值。程序一打开计算机屏幕就显示 SPSS 程序的主窗口,主窗口显示一个数据文件,屏幕的最上面是工具菜单栏和一些工具图标,最左面是实体的标号(见表 13-1)。

数据可以手工输入,也可以从文本文件,EXCEL 等电子表格或数据库文件中整体地转换引入。文件转换的操作过程如下:打开 SPSS 程序后,单击"File",从下挂菜单中单击"Open"→Data。这时会出现一个对话框,选择路径和文件类型,可以找到需要转换的文件。确认后单击"打开",就执行文件的转换,SPSS 数据文件建立完成。如果从 Excel 表格转换,Excel 表格第一行的内容自动转换为各变量的名称。文件转换时需要注意变量类型的一致性。

主窗口显示的数据文件的左下角有一个"Data View/Variable View"切换开关,可以从显示数据切换到显示变量。在显示变量的状态时,可以观察到各变量的类型并可进行编辑,包括选择和改变变量的类型、显示宽度、小数点位数和定义变量的标识符等。变量可以是数值型、字符型或日期、dollar 等其他类型。数据文件建立后,可以以"sav"为扩展名作为 SPSS 数据文件保存,也可以保存为其他格式的数据文件,为别的软件使用。

SPSS 数据文件的界面是英文的,但在 Windows XP 等较高的版本中,变量名、变量标识符和名称变量的取值等都可以输入汉字。

(二) 数据文件的预处理。

单击工具栏中的"Data",通过下挂菜单,可以插入或删除实体、插入或删除变量、定义或改变变量、对实体进行选择、排序和加权以及数据文件的分解或合并等,多数操作方式与 EXCEL 等电子表格相似,无需详细介绍,这里仅对实体的加权和选择稍加说明。

1. 实体的加权。在分析数据的交叉列联表时往往需要对实体加权。本书的第十章和第十一章中讨论名称变量和有序变量的关联和相关时,其基础数据是统计记录交叉分类频次值的 $r \times c$ 列联表,如表 10-11a。

墓式 \ 年龄段	青少年 Y	壮年 M	老年 O
简单土坑 T	23	19	11
木制墓室 M	12	17	13
石砌墓室 S	10	16	15

这是一张研究墓葬类型和墓主人年龄段两个变量间关联的数据表,表中单元格的内容是相应的频次值。使用 SPSS 软件处理这类列联表,要使用数据加权的命令。实际操作过程如下:

第一步是将列联表 10-11a 的内容,按通常 SPSS 的格式建立数据文件,如表 13-1 所示。

表 13-1 SPSS 数据文件的格式

	墓式	年龄段	频次
1	1T	1Y	23
2	2M	1Y	12
3	3S	1Y	10
4	1T	2M	19
5	2M	2M	17

续表

	墓式	年龄段	频次
6	3S	2M	16
7	1T	3O	11
8	2M	3O	13
9	3S	3O	15

表中共有 9 个实体,每个实体有 3 个变量。它们是墓式,墓主人的年龄段和该类实体,即墓葬的频次值,表的第 2,3 两列反映了墓葬的 3 种墓室和墓主人 3 个年龄段的 9 种交叉情况,对应于表 11-10 的 9 个单元格。这两列属名称变量,各有 3 个赋值,第 4 列单元格的内容是频次,它是数值变量。

建立表 13-1 后,第二步是对实体加权。为此单击"data"→"weight cases",在出现的对话框中选择"频次"作为加权变量。按"OK"就完成了对每个实体以其频次值为权的操作。这样就可以对表 11-10 的数据作 χ^2 检验,计算各种关联强度等。需要说明,在加权操作中,权的数值中不能出现"0"。如果实际的频次值中包含"0"值,需要用一个小的数值(例如 0.001)取代 ,否则加权操作会被停止。

2. 实体的选择。对 SPSS 数据文件中的实体可以按一定的标准进行选择,其操作程序如下。单击"data"→ "select cases", 然后在出现的对话框中输入实体选择的条件,例如输入条件是"墓室 = "1T"and 墓室 = "2M""(因为 1T 和 2M 为名称变量的值,其前后要用""号括起来),那么土坑墓和木制墓被选。也可以在数据文件中专门建立一个分组变量,然后根据分组变量的数值选择实体。执行实体选择操作后,在数据文件中将自动产生一个新的变量"filter _ $ ",该新变量对于被选实体赋值为 1,未选实体赋值为 0。还应注意,实体一旦被选或加权,那么它们的被选或加权的状态将被保留,除非以后用户改变权重或重选。

13.2　数据的转换

单击主窗口最上面工具栏的"Transform",将下挂另一个菜单,包含数据转换的各种命令。

1. "Compute"命令是通过数学运算产生一个新的变量。在"Compute"对话框中(1)键入新的目标变量的名称,定义目标变量的类型和标识符等属性。(2)按"if"键,输入条件以选择需要对新变量赋值的实体。(3)在"Numeric Expression"或"String Expression"框中键入希望生成的新变量的表达式。书写表达式时,除使用字母,数字和各种数学运算符外,还可使用已打开的数据文件中的变量和 SPSS 软件的内部函数。SPSS 有 70 多个内部函数,包括数学函数,统计函数,分布函数,字符串函数等等。

2. "Count"命令是统计每个实体在变量表中同类值出现的次数,并将统计结果生成一个新的数值变量写入已打开的数据文件。操作过程是在"Count"对话框中(1)键入新的目标变量的名称和属性。(2)将需要统计的所有变量转移到变量框。(3)按"define values"键,在其对话框中定义所需统计的"值",它可以是数值变量的一个数值,几个数值或数值

范围,也可以是字符串。定义完成后,按"Add"键确认。这个操作可以多次重复。完成"define values"操作后,按"Continiu"键,回到"Count"对话框,按"OK"键,命令即被执行。(4)如果仅需对部分实体执行 Count 命令,则在按"OK"键前,先按"if"键,输入相应的选择条件,可选择需要统计的实体。

3."Record"命令对变量的取值作改变,或称重编码。改变后的新结果可以存入同一个变量,即进行改写操作,也可以作为一个新变量写入数据文件。(1)如果选择对原变量重编码,则在对话框中先选择要重编码的变量,再按"Old and New Values"键。在新出现的对话框中分别键入需改写的旧值和修改后的新值,按"Add"键转移确认。这个操作可重复进行,按"Continiu"键,回到"Record"对话框,按"OK"键,命令即被执行。需改写的旧值可以是单个数值,也可以是数值范围,还可以是缺失值。(2)如果重编码后作为一个新变量写入文件,则在对话框中选择要重编码的变量后,键入新的目标变量的名称和它的标识符等,按"Change"键确认。再按"Old and New Values"键,往下的操作同前。"Record"对话框中的"if"键的功能与"Compute"命令中相同。

4."Rank"命令是将实体按某个或几个变量取值的大小排序,并将排序结果作为新变量(秩变量)写入数据文件。新变量的名称是依据原来变量的名称自动生成的,如果原变量名是 Var001,那么新生成的秩变量的名称为 rVar001。操作过程:在"Rank"对话框中(1)输入原变量名(2)选择排序的方法(一般选"Rank")和(3)确定对原变量取值相等的实体的秩赋值的方法,程序即可执行。如果需要对实体作分组的各自排序,应将分组变量输入"By"栏。

在上面各项命令的对话框中都有一个"Paste"图表按钮,它将打开一个"SPSS Syntax Editer"对话框,显示所要执行过程的 SPSS 源程序,也可在此框内,对程序进行编辑,编辑完后,点击"Run"执行。

13.3 基本统计分析程序

单击主窗口上面工具栏中的"Analyze",屏幕将显示列出所有分析程序的下挂菜单。本节仅选择菜单中与本书内容有关的程序作简单说明。

(一)"Descriptive Statistics"程序组。

点击"Descriptive Statistics" 将出现下一层菜单,列出描述性统计分析的基本命令。

1."Frequancy"命令是显示实体的分布。在"Frequancy"对话框中首先要选择一个变量。该命令的执行将产生一个全部实体按照所选定变量的分布表,该表将显示各类实体的频次,频率,并对实体排序后显示累积频率。

在对话框中单击"Chart", 那么执行结果还将显示分布图,用户可选择直方图,圆瓣图,长条图等,如选择直方图,图上可同时叠加拟合的正态分布曲线。

在对话框中还有一个"Statisrics"开关,如开启,命令执行结果中还将按用户要求给出数据的集中量数和差异量数。

SPSS"Analyze" 菜单下所有的分析程序执行后,都会把执行结果写入一个"Output - SPSS Viewer"文件中,其内容可以以扩展名为"spo"的文件保存,也可以将其内容拷贝到

"Word"等软件中。"Output - SPSS Viewer"文件中的图形,可以利用SPSS软件提供的图形编辑器(SPSS Chart Editor)进行编辑,并用多种图片格式单独存储,从而方便地将SPSS执行结果与其他软件连接。

2. "Descriptive"命令给出最基本的描述性统计的结果,包括平均值,标准差等。在对话框中有一个"Option"按钮,通过它还可以选择计算"最大值、最小值、方差、峭度、偏斜度等统计量。"Descriptive"中的很多子命令包含在"Explore"命令中,后者的功能更强,建议使用"Explore"命令。不过"Descriptive"命令有一个特殊的功能,它能对原始数据根据其标准差标准化处理,并把标准化后的Z分量作为新变量写入数据文件中。

3. "Explore"命令将以表格的形式给出全部描述性统计的结果。打开对话框后,把对实体组所要分析的变量名输入到对话框的"Dependent List"栏中,选择所要计算和输出的内容后,单击"OK"就完成操作。表13-2是对25件东周青铜钟和鼎的锡百分含量的描述性统计的输出结果。原始数据为:(16.73,18.10,17.50,19.66,13.72,12.40,18.21,12.62,16.93,13.90,15.30,12.00,15.30,15.20,12.63,15.90,12.49,13.44,13.76,14.46,14.60,14.12,15.31,17.72,17.45)%。

表13-2 对25件东周青铜钟和鼎中锡的百分含量执行"Explore"命令后的部分输出结果

		Statistic	Std. Error
Mean		15.1780	0.42763
95% Confidence Interval for Mean	Lower Bound	14.2954	
	Upper Bound	16.0606	
5% Trimmed Mean		15.1172	
Median		15.2000	
Variance		4.572	
Std. Deviation		2.13814	
Minimum		12.00	
Maximum		19.66	
Range		7.66	
Interquartile Range		3.6100	
Skewness		0.332	0.464
Kurtosis		-0.868	0.902

表的最后一列给出平均值、偏斜度和峭度的标准差。"Explore"命令还能给出数组的各分位数的数值,如表13-3所示。表中"50%"位的分位数值也就是中数,其左右两个数值是上、下四分位数。

表13-3 "Explore"程序给出25件东周青铜钟和鼎的锡百分含量的各分位数

5	10	25	50	75	90	95
12.1200	12.4540	13.5800	15.2000	17.1900	18.1440	19.2250

"Explore"命令除给出上面的描述性统计分析结果外,用户还可以要求它绘制出实体分布的直方图、茎叶图和箱点图。"Explore"命令可以对几个变量同时进行分析,也可以把

实体分组分析，只需将分组变量输入对话框的“Factor”栏。分组分析时，各组的箱点图输出在同一张图上，十分便利于各组数据间的中数和四分位差之间的比较。

4.“Crosstabs”命令应用于名称变量频次列联表的关联研究和有序变量的相关研究。执行本命令前应先按 13.1 中所述的给各实体按其频次加权。单击“Crosstabs”命令，在对话框中输入行变量和列变量的名称，点击“OK”键，程序对名称变量给出 χ^2，ϕ，V，λ 和 Goodman and Krustal's τ 等关联强度系数，而对有序变量给出 Gamma 和 Kendall' τ_b 和 τ_c 等等级相关系数，同时给出相应的显著性水平。

（二）“Compare means”程序组。

选“Compare means” 将出现下一层菜单，均为执行总体平均值间比较的统计分析命令。

1. 单击“Mean”命令出现一对话框，将需要求平均值的变量输入到“Dependent list”栏，同时必须将分组变量输入到“Independent list”栏。即使只有一组实体，也需要输入分组变量，这时只要将对全部实体的分组变量赋以同一个数值就可以。对话框中的“Option” 钮提供各种描述性参数的选择，命令的执行给出各组的平均值和其他描述性参数。“Mean”命令还可以对各组数据作 ANOVA 分析，给出组间离差平方和和总离差平方和的比值，称为 Eta 系数。

2. “One sample T test”命令执行单总体平均值的假设检验。将变量名和总体的平均值分别输入对话框后（后者输入到“Level”栏），命令即可执行。这里的“Option”钮提供检验置信度的选择和处理数据文件中缺失值方法的选择。

3. “Independent-samples T test” 命令执行两总体平均值一致性的假设检验。输入分析变量名，分组变量名和分组标准后即可执行。对话框中“Option”钮的作用与“One sample T test”命令情况下相同。输出结果中包括在方差一致性检验通过和不通过两种情况下的两总体平均值一致性的检验结果和两总体平均值之差的区间估计。但是这个程序并不对两总体方差的一致性作检验。

4. “Paired-samples T test”命令执行成对样本的总体平均值一致性的假设检验。需要同时输入一对变量。其他方面与前述内容相同。

5. “One-way ANOVA”命令执行一元方差分析。输入分析变量和分组变量后，命令即可执行并输出 ANOVA 分析表。“Option”钮可提供多种选择，例如可要求输出描述性统计的各参数值，进行方差的一致性检验等。如果“One-way ANOVA”检验拒绝了“各组平均值无显著差别”的原假设后，“Post-Hoc”键提供进行各组两两之间比较。这里有多种方法可供选择，例如在通过方差一致性检验的条件下选择“LSD”或“Tukey”方法，执行结果将显示各组两两间平均值一致性的检验结果和哪几个组之间能通过平均值一致性检验等

（三）“Correlerate”程序组。

本程序组包含计算各种相关系数和距离系数的程序，由三组程序组成。

1. “Bivariate”程序计算两个变量之间的简单相关系数，包括皮尔逊相关系数，斯皮尔曼相关系数和 Kendall's τ_b 系数，并进行双测和单测的检验。

2. “Partial”程序计算偏相关系数，需要将分析变量和控制变量分别输入对话框。

3. “Distances”计算各种距离系数。在对话框中可选择需要计算实体间的还是变量间

的距离，选择表示距离的是相似系数还是相异系数。对话框中的“measure”按钮对数值变量，二元变量和频次变量分别给出不同的距离系数供用户选择，例如对数值变量可以选择计算并输出欧氏距离、城市街道距离或车贝舍夫距离等等。

（四）“Regression”程序组。

本程序组的功能是给出回归方程，检验回归方程的稳定性和进行残差分析等。可进行线性的和非线性的回归，一元的和多元的回归，是一个有多种功能的程序组。与本书内容有关的仅是一元线性回归分析，为此在“Linear”对话框中分别输入一个自变量和一个应变量后就可执行，输出相关系数，回归参数和相应的显著性检验结果。这里再次说明这是一个功能很强的程序组，一元线形回归仅是其很小的部分。

（五）“Classification”程序组的功能是对实体分类。

本书将在第十四和十五章，结合介绍各种多元的分类和归类方法的原理和应用时详细介绍 SPSS“Classification”程序组的功能和操作，这里仅作简要说明。“Classification”程序组包含 3 个次级程序。

1．“K-means Cluster Analysis”程序对实体进行快速的非等级的分组。在对话框中输入分类中所考虑的变量，确定需要分几组和计算过程中的迭代次数后程序即可执行。各分析变量在各组中的初始中心值可由用户给定，也可以由计算机生成。本程序的应用简单方便，但要求某些先决条件。也许应同时使用“Hierarchical Cluster Analysis”方法作对比。

2．“Hierarchical Cluster Analysis”程序称为系统聚类程序。在对话框中输入分类中所考虑的分析变量。按“Method”键后有 3 项重要的选择，(1)选择原始数据标准化的方法，(2)选择使用哪种距离系数作为实体间相异程度的度量和(3)选择系统聚类的方法。程序输出的格式，包括树支状图，冰柱图等。如在“Statictics”对话框选“Agglomeration Schedule”可显示逐步聚类的过程，选“Proximities”可显示聚类过程所基于的相异系数矩阵，它是由距离系数度量方法的选择所决定的。详细情况见 14.4 节。

3．“Discriminant Analysis”程序进行判别分析。在主要对话框中需要输入分析变量（即自变量）和分组变量后就能执行。但程序执行前另有若干选择项，包括自变量进入的方式，以及“Classify”对话框中的先验概率，缺失值的处理，是否要求执行“Leave-one-out-classification”以及输出表格和图形的内容和形式等。在“Statistics”对话框中可以要求计算和显示相关系数，各种协方差，各变量各组的平均值，非标准化情况下判别方程的系数和执行一元方差分析等，详细的情况将在第十五章中结合应用实例加以说明。

（六）“Data Reduction”程序是进行主因子分析。关于其功能和操作过程将在第十六章详细讨论，这里仅是简要介绍。在主要对话框中输入分析变量后程序即可执行。但主因子分析需要根据分析的目的和原始数据结构先确定一系列选择项。

1．在“Discriptive”对话框中，用户可要求作单变量的描述性统计和计算和输出各种相关系数矩阵。其中的 KMO and Bartlett's Test-of-Sphericity 统计量显示整套数据是否适宜于主因子分析，Anti-Image 反象相关系数矩阵显示每个变量对于当前因子分析的适宜性度量，建议选择。

2．“Extraction”对话框中有几项重要的选择。(1)“Method”栏中提供多种提取因子的方法，它们的差别是拟合优度的定义标准不同。默认的方法是选择提取“Principal Compo-

nent”,即主成分分析,这也是考古学定量分析中最常用的方法。最大似然法有时也被选用,不同因子提取方法给出结果间的比较可显示主因子分析的结果是否稳定。(2)“Analyze”栏中要求用户在相关系数矩阵和协方差矩阵之间作选择,不同的选项会导致有一定差异的分析结果。建议选程序默认的相关系数矩阵。在这个对话框中(3)还可以选择提取特征值的标准和(4)选择某些显示方式等。

3. “Rotation”对话框提供用户选择,是否需要转动主因子轴和转动的方式,转动的目的是简化数据结构以便更清楚地显示主因子与原始变量间的关系。

4. “Scores”对话窗口应选择“Save as Variables”,将实体的各主因子或主成分得分值作为新变量记录于原始数据文件中。此外还应选择显示因子得分矩阵,以了解原始变量对各主因子的贡献。

5. “Option”按钮选择缺失值的处理方案和要求因子得分的输出按大小排列等。

(七)“Nonparametric Test”非参数假设检验程序组提供单样本,独立和相关样本的非参数假设检验。

1. “Chi-square”是对单样本进行 χ^2 检验。在对话框的“Test Variable List”栏输入分析变量,在“Expected Values”栏中默认的是“All catagories equal”,各期望值相等,也可以在“Value”栏中依次输入与每个实体取值相对应的期望值,例如根据均匀分布假设前提下计算得到的期望值。程序执行后将输出各组实际观察值和期望值的频次表,χ^2 值,自由度和相应的显著性水平。单击“Option”钮,还可显示分析变量的描述性参数和各分位数值。

2. “Binomial”是对单样本的比例数作二项分布检验,在对话框中输入分析变量名和期望比例数 p 后,程序即可执行。要注意的是,输入的期望比例数是对应二元分析变量的第一个取值。另外如果分析变量不是二元变量,它的取值数大于 2 时,则要在“Difine Dichoyomy”的“Cut point”中输入一个分割值。小于或等于分割值的数据归入第一组,大于分割值的数据为第二组,将原始分析变量转化为二元变量,然后才能进行二项式分布检验。

3. “Run”是游程检验程序,本书未讨论游程检验问题。

4. “1 Sample K-S”执行单样本的 Kolgomorov-Smirnov 检验,可检验样本是否符合正态,均匀,泊松或指数分布,但分布的参数不能选择。

5. “2 Independent Samples”执行两个独立样本的非参数检验。在对话框中分别输入分析变量和分类变量后,再选择检验的方法,程序即可执行。检验方法中包括本书前面已介绍的基于秩和的“Mann-Whitney”方法和基于百分累加曲线的“Kolgomorov-Smirnov”方法。对于前者输出结果表给出近似的和精确的两个显著性水平值,可根据样本的实体数大于或小于 40,分别选精确值或近似值。

6. “2 related Samples” 执行两个相关样本或成对样本的非参数检验。在对话框中输入变量对和选择检验方法后,程序即可执行。检验方法有本书第七章曾介绍或提到的符号检验和 Wilcoxon 符号秩检验。

13.4 绘图程序

SPSS 软件能执行一定程度的绘图功能,可以在“Graphs”菜单下执行。另外上节介绍

的很多统计分析程序的执行自动生成(或选择"Plot"要求生成)各种显示图。单击这些图,即可进入SPSS的图形编辑程序"SPSS Chart Editor"窗口。在图形编辑窗口,对图中的点,线等图形元素的形状,大小,粗细,填充,颜色,对坐标轴的单位,分格,名称等进行编辑。编辑后的图可以以多种格式输出,也可直接粘贴到"Word"文件中。

Graphs菜单中还有P-P和Q-Q两个命令,可用于粗略考察数据是否接近正态分布,详见7.5节。

13.5 在线帮助

SPSS提供多种帮助方式:(1)在SPSS主窗口上端的工具栏中,单击"Help",可出现帮助菜单。可选择按"目录"(程序功能)或按"索引"(字母次序)寻找帮助。帮助内容除对程序,方法的基本原理作说明外,还告知用户怎样操作(How to)。(2)在很多的程序命令窗口中都有"help"键,对当前的命令作解释和提供帮助。(3)在分析程序执行结果的输出文件"Output-SPSS Viewer"中也能寻求帮助。将鼠标点在有关条目或表格上,单击鼠标右键也能得到相应的帮助,可以打开"Result Coach"窗口,程序会对分析的结果进行举例解释,这非常有助于对统计学知识准备不足的用户。(4)此外SPSS软件还提供"Tutorial"和"Statistic Coach"等专门的教学程序,供初学者学习使用。



下　　篇

多元统计方法在考古研究中的应用



本书上篇讨论了单变量和两个变量的情况，讨论了实体对于单变量和双变量的分布、单变量实体组的集中量数和差异量数、两个变量之间的相关关系，介绍了统计推断的基本思想和方法等。但是在很多情况下实体具有多方面的属性，仅用一、二个变量来描述是不充分的。例如在墓葬分期中，需要考虑墓葬中多种器物的存在和数量，在陶瓷的产地溯源中要比较陶瓷中二三十种元素的含量，在古人类颅骨的种族判别中要分析比较描述头骨形态特征的几十种观测量。这些考古研究课题都涉及大量的实体，而每个实体又被多个变量所描述，原始数据复杂庞大。考古学家为了从庞大烦琐的数据结构中寻求其内涵的关系和规律，需要凭自己的经验，花费大量的精力和时间。但是传统的研究方法难免有疏漏之处，其研究结论还可能隐含研究者个人的观点和倾向。为了从庞大复杂的数据中找出其内涵的关系和规律，特别是对多变量实体进行分类和排序，研究者发展了各种多元统计方法。对于各种多元统计方法，其计算规则的共同特点是简化数据结构，在简化了的数据结构中更容易观察，发现原始数据中所包含的关系和规律。例如本篇第十六章要介绍的主成分分析方法，它是在保留原始数据中绝大部分信息的前提下，将原始数据简化到仅有二、三个综合变量的数据，然后在降维后的二、三维的空间中对实体进行分类排序。这种计算过程称为数据的降维。第十五章的判别分析则是计算多变量实体的少数几个判别函数值，根据少数几个判别函数值就可以对实体作归类。第十四章的聚类分析和第十七章的 Brainerd-Robinson 排序方法，则是在多维变量的空间中计算实体两两间的相似系数，然后根据相似系数矩阵对实体进行聚类或排序。当然，主成分分析、判别分析和聚类分析等多元分析方法在简化数据结构的过程中也涉及某些项目的选择。对于同一组原始数据，不同的选择可能会得到不尽相同的结果，也就是说，最后的分析结论会因不同的选项而带有研究者的主观因素。但是这些主观因素是“公开的”，其他人也能看到的。多元分析是由计算机帮助实现的，计算分析过程快速，可以通过观察改变选项对分析结论的影响，帮助揭示变量间的关系和评估分析结论的可信度。

多元统计方法也称为多变量分析方法。之所以称为统计方法，是因为这些计算方法往往是基于平均值、方差、相关系数等概率统计学的基本概念。但是这些方法所处理的数据并不是统计意义下的样本，一般不要求随机抽样，研究结论也不被要求外推，不涉及显著性检验的问题。例如对某个墓地的墓葬分期，对一批陶瓷片按其化学组成分类或古人类头骨按其形态分类等，对所处理的样本可以看成是总体。这与本书上篇中所进行的统计推断是不同的。

多元统计方法作为数学学科的一个分支，创建于 20 世纪之初。各种多元分析方法都涉及大量的计算，因此它们的发展和应用的普及是与计算机的普及分不开的。二次大战以后多元统计方法迅速地发展，特别是在生命科学、农业科学、经济学和社会学等学科得到极为广泛的应用。有人认为，数学特别是多元统计方法的应用使生命科学获得了第

二次生命，使经济学走上计量的道路从而成为真正意义上的科学。多元统计方法在考古研究中的应用也正在逐步开展并有光明的前景。

鉴于基础统计和多元统计的计算过程都借助于计算机软件，上篇的第十三章介绍了SPSS统计软件最基础的知识和使用方法，本篇介绍主成分分析、判别分析和聚类分析等多元分析方法时也是结合SPSS软件的使用进行的。这将有利于读者学以致用，应用这些方法于实际的考古资料的分析研究。

第十四章　实体的分类和等级聚类分析

14.1　数量分类方法一般介绍

实体的分类是指把一个实体群按其性状特征分成若干组，并使得每个组内实体间的相异性尽量小，而不同组的实体间的相异性尽量大。分类是人类认识自然的一种基本方法，是从大量繁杂的资料数据中寻找关系和规律的方法。分类研究在生物学的发展中起了关键的作用，生物分类学将物种作为分类的基本单位，并建立起了种、属、科、目、纲、门和界等自小到大的类群，反映分类阶元的梯级结构，并一定程度上反映种系的发生和进化。在考古学中器物的分型分式是器物层次的分类，用以研究器物随时代的演化和器物的地区性特性；而文化类型的划分则是更高层次的分类，它是在一定地域建立考古学文化谱系演化的基础。

分类的方法很多种，但可以分成等级分类方法和非等级分类方法两大类。等级分类方法又分为等级聚类和等级分划两种。等级聚类也称系统聚类，它是在一个包含 n 个实体的实体群中先将 2 个性状最相近的实体聚合，并且看作一个新的实体，再在 $(n-1)$ 实体中找出 2 个性状最相近的实体聚合，经过 $(n-1)$ 次这样的聚合，得到包含全部 n 个实体的聚合。因为聚合有先后，最终得到一个自上而下等级状的聚类结果，称为树枝状聚类图。等级分划是根据一定准则将全部 n 个实体划分成 2 组，然后再对其中的一组一分为二，这样重复进行，直到可以认为已分划的各子组内的实体已是同质的，不需再分划。也就是说等级分划有自己的终止规则。等级分划最后将给出一个由下而上有等级的树枝状分类图。

与等级分类相对应的是非等级的分类方法。譬如说，将全部实体同时分为预定的若干组，然后根据规定的标准用迭代方法对各组成员进行调整，最后得到一个网状结构的分类图。非等级方法由于计算工作量大和别的一些困难，其发展和应用的普遍性不及等级分类方法。非等级分类的结果，组内实体的同质性好；而等级分类的树枝状图能多少反映实体间的"谱系"关系，但这是以牺牲同组实体的同质性为代价的。等级聚类的另一个问题是，如果因某种原因某个实体在聚类的早期阶段被不适当的聚合，就可能对后面的聚类过程产生不良的影响。

第十五章将讨论的判别分析，严格地说不属于分类方法，而是一种归类的方法。这里外在的因素已能确定实体应该分成几类，即实体的类属关系是已知的。譬如说有一批瓷片，已经知道它们来自几个不同的产地。在这个先验的条件下，又根据实体本身内在的性状，譬如说根据这批瓷片的化学元素组成分类，观察按照化学组成分类的结果与已知的按照产地的分类是否符合。此外判别分析还能将未知归属的实体，例如未知产地的瓷片，归到已知的合适类别。人工神经网络方法基本上也是一种归类的方法。

第十七章将讨论的主成分分析，其分析结果往往用实体在由前二三个主成分为坐标系的空间中的分布来表述，实体的分布情况可以为它们的分类或排序提供重要信息。因此主成分分析同时可看成分类方法和排序方法。

上述的分类方法多数是基于实体的可测量性状的分类方法，都属于数量分类方法。在具体讨论各种分类方法之前，14.2 和 14.3 两节先介绍数据的转换和相似系数，它们是执行多种分类方法前的数据准备工作。然后再依照等级聚类，等级分划和非等级分类的次序进行讨论。

14.2 原始数据的转换

聚类分析的基础数据是原始数据表。假设有 n 个实体，每个实体用 m 个属性来描述，原始数据表如下式(14-1)所示：

$$(x_{ij})_{n\times m} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \tag{14-1}$$

这是一种矩阵的表达形式。矩阵有 n 行，m 列，共 $n \times m$ 个元素组成。每一行代表一个实体，可以看成一个 m 维的矢量；而每一列代表一个变量在 n 个实体中的取值，可以看成一个转置了的 n 维的矢量。所谓转置就是行(列)通过 90 度的旋转转化为列(行)。

这里作一个说明，多元统计分析的理论和计算过程都涉及矩阵代数。本书基本内容的阅读和学习并不要求读者掌握有关矩阵的知识，但在某些章节使用了一些矩阵代数的术语。我们会对所用的术语作必要的说明，将不致引起未掌握矩阵基本知识的读者的阅读困难。

聚类分析可以对实体进行聚类，也可以对变量进行聚类。前者称为 Q 型聚类，往往用实体之间的“距离”作为实体间相异程度的指标；后者称为 R 型聚类，往往用变量之间的相关系数作为变量间相似程度的指标。在 14.1 节中曾指出，聚类分析的基本过程是把两个性状最相近的实体聚合为一类，因此定义和计算表征实体间性状相近程度的各种相似系数是执行聚类过程的前提。14.3 节将专门讨论各种相似系数的定义。

原始数据中的各个变量往往使用的是不同的测量单位，即使使用同样的测量单位，不同变量取值的变化范围也有差别，这将影响相似系数的计算结果。举例来说，有一群人，每个人有 不同的身高(x) 和体重(y)，在以身高和体重为坐标的图上，每个点代表一个个体。点与点之间的距离是 $d = \sqrt{x^2 + y^2}$，它反映个体间身高和体重的差别，距离越大，个体间的差异也越显著，因此距离 d 可以作为人与人之间的相异系数，或称距离系数。相异系数和相似系数是互补的，都是表征实体间关系亲疏的度量。但是距离 d 的数值大小依赖于身高和体重的测量单位。如果人的身高用厘米，体重用市斤作为测量单位，那么 x 与 y 的数值大小很接近，都是一百几十，身高和体重对 d 的贡献是差不多的。但如果改用米作为身高的测量单位，那么在数值上 $y >> x$，这样点之间的距离，也就是表征人与人之间亲疏程度的量，将主要由体重来决定，身高将几乎不起作用。反之，若身高用毫米

作测量单位,体重用公斤表示,那么在 x 与 y 为坐标的图上,点之间的距离将主要由身高来决定,体重将几乎不起作用。表征实体间亲疏程度的量依赖于原始数据测量单位的情况当然是不能被接受的。为此要对原始数据作某种变换,变换的目的是使得实体间相似系数的计算不因原始数据使用不同的测量单位而受影响,并且使得各个变量对相似系数有大致相等的贡献。数据的这种转换称为数据的标准化。

数据的标准化有多种方法可以实现,下面介绍几种常用的数据转换方法。

(1) 数据的中心化。数据中心化的过程如下。

先计算每个变量的平均值,即对表 14-1 中每一列的数据计算其平均值 $\bar{x}_j$,然后该列的每个数据均减去它们的平均值 $\bar{x}_j$

$$x'_{ij} = x_{ij} - \bar{x}_j \tag{14-2}$$

中心化后,每个变量的平均值都为 0,但是它们间的方差还是有差异的。

(2) 数据中心化后再用标准差进行标准化。

计算每一列的数据的标准差 s_j,然后将公式(14-2) 除以 s_j,进一步转换数据。因此用标准差进行标准化的公式为

$$y'_{ij} = \frac{x'_{ij} - \bar{x}_j}{s_j} \tag{14-3}$$

这个数据转换过程实际上就是第四章公式(4-28) 计算 Z 分量$\left(Z = \frac{x-\mu}{\sigma}\right)$的过程。因此式(14-3) 也可写成

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{14-4}$$

变换后的数据矩阵为

$$(Z_{ij})_{n\times m} = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{s_1} & \frac{x_{12}-\bar{x}_2}{s_2} & \cdots & \frac{x_{1m}-\bar{x}_m}{s_m} \\ \frac{x_{21}-\bar{x}_1}{s_1} & \frac{x_{22}-\bar{x}_2}{s_2} & \cdots & \frac{x_{2m}-\bar{x}_m}{s_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{x_{n1}-\bar{x}_1}{s_1} & \frac{x_{n2}-\bar{x}_2}{s_2} & \cdots & \frac{x_{nm}-\bar{x}_m}{s_m} \end{pmatrix} \tag{14-5}$$

经上面的转换后,每个变量不仅平均值均为 0,而且它们的方差相等,均为 1。如果原始数据基本服从正态分布,那么矩阵(14-5)的元素 Z_{ij} 中,约有 95%取值在 −2 与 2 之间,99.7%取值在 −3 与 3 之间变动,式(14-3)的数据标准化过程又称为数据的正规化,是最常用的数据标准化的转换方法。

(3) 用极差进行标准化。

在每列数据中找出最大值和最小值,两者之间的差值就是该列数据的极差 R_j 。

$$x'_{ij} = \frac{x_{ij} - j\text{列的最小值}}{R_j} \tag{14-6}$$

式(14-4)是用极差进行数据标准化的公式。用极差标准化后,每列数据中的最大值为 1,

最小值为 0,其他数据在 0 与 1 之间,接近了数据标准化的目的。

但是每列数据的最大、最小值都是该组数据的极值,有可能是偏离平均值和中数甚远的歧离值。如果某列数据的最大值偏离中数很远,该组数据按极差标准化后,大多数数据值将靠近 0 而偏小。以后在计算相似系数时,该变量的贡献某种程度上会被压低。这是用极差进行标准化的缺点。

(4) 用总和进行标准化。

首先计算各列数据的总和 $\sum_i x_{ij} = s_j$。

$$x'_{ij} = \frac{x_{ij}}{s_j} \tag{14-7}$$

式(14-7)是用各列数据的总和进行数据标准化的公式。标准化后全部数据均是小于 1 的正值,各列的和均为 1,即 $\sum_i x'_{ij} = 1(j = 1,2,\cdots m)$。

数据的标准化还有其他的方法,例如用各列数据的最大值标准化,用各列数据的平方和(称为模)标准化,用各列数据的离差平方和的开方来标准化等。

前面的讨论都是对变量进行数据的标准化。同样可以对实体进行数据的标准化,这需要计算各行数据的平均值、标准差、极差、总和和最大值等。还可以同时对实体和变量标准化。选择对实体还是对变量进行标准化,选择哪种标准化的方法取决于实际的研究问题。但是对变量用 Z 分量的方法标准化,即正规化是最常用的方法。

14.3　实体间的相似系数

聚类分析的基本过程是把全部实体或者变量,根据它们之间的相似程度逐步聚合为一类。因此定义和计算表征实体间或者变量间相近程度的各种相似系数是执行聚类过程的前提。相似系数的种类有多,它们适用于不同的数据类型,而且也因为是对实体还是对变量进行聚类而不同。我们将介绍三种不同类型的相似系数。

14.3.1　距离系数

当对实体聚类,而且描述实体的变量都是数值变量时,一般用距离系数来表征实体间的相似程度。在用变量作为坐标的空间中,每个实体可以看作为空间中的一个点。两个点之间的距离 d_{ik} 反映这两个点所代表的两个实体间的相异程度。d_{ik} 越小,表明实体间的性状越接近;d_{ik} 越大,表明实体间的性状差异也越大。因此,距离或距离系数实际上是实体间的相异系数。前面我们已经说明相异系数和相似系数是互补的,都是表征实体间关系亲疏的度量。它们之间的转换是很容易的。例如我们有一组相异系数,它们在最小值 0 和某个最大值 L 间变动,现在用 L 值去减每个相异系数,原来相异系数的最小值 0 转化为 L,而原来相异系数的最大值 L 就转化为 0,即简单的减法运算把一组相异系数转化为一组相似系数了。因此,下面我们将不再着意区分两者间的差别,而且统称为相似系数。很多计算机软件直接处理相异系数,而不进行它们间的转换。

距离系数也有不同的定义,这里介绍三种距离系数。

(1) 绝对值距离，也称城市街道距离：

$$d_{ik} = \sum_{j=1}^{m} | x_{ij} - x_{kj} | \tag{14-8}$$

(2) 欧氏距离和欧氏距离的平方：

$$d_{ik} = \left(\sum_{j=1}^{m} (x_{ij} - x_{kj})^2 \right)^{\frac{1}{2}} \tag{14-9}$$

欧氏距离是最常用的距离系数。有时也用欧氏距离的平方 $d_{ik}^2 = \sum_{j=1}^{m} (x_{ij} - x_{kj})^2$ 作为相似性度量，后者具有数据可加性的优点。

*(3) 马氏(Mahalanobis)距离，这是多元统计分析中一个十分重要的距离系数。其定义和计算公式如下：

$$d_{ik} = [(\boldsymbol{x}_i - \boldsymbol{x}_k)'\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_k)]^{\frac{1}{2}} \tag{14-10}$$

式中 $\boldsymbol{x}_i$ 是代表第 i 个实体的矢量，也可以看作($m \times 1$)矩阵，($\boldsymbol{x}_i - \boldsymbol{x}_k$)是 2 个矢量的差，$\boldsymbol{S}^{-1}$ 是实体内积系数矩阵 $\boldsymbol{S}$ 的逆矩阵，为($m \times m$)矩阵，而矢量$(\boldsymbol{x}_i - \boldsymbol{x}_k)'$是矢量$(\boldsymbol{x}_i - \boldsymbol{x}_k)$的转置，式(14-10)是这 3 个矩阵的乘积(矢量也可以看成矩阵，只是仅有 1 列或仅有 1 行)。关于矩阵的运算，我们不可能作详细的讨论。我们所以在这里提到马氏距离，是因为将来在学习判别分析时会用到这个概念。此外如果实体间有两个完全相关的变量，虽然这两个变量转换前的取值是不相等的(固定的倍数)，但在数据正规化后，这两个变量的取值将相等，他们对实体间的欧氏距离将没有贡献。但马氏距离却不受变量全相关的影响。当变量间完全不相关时，$\boldsymbol{S}^{-1}$ 将是一个单位矩阵，公式(14-10)简化为公式(14-9)，马氏距离简化为欧氏距离。因此马氏距离又称广义距离，马氏距离在多元统计分析中起重要的作用。

所有距离系数都是大于或等于 0 的。对所有的实体两两间计算了距离系数后，可以写成一个 n 行 n 列的矩阵，称为距离系数矩阵。

$$(d_{ik})_{n \times n} = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix} \tag{14-11}$$

主对角线上的元素 d_{ii} 都等于 0，因为每个客体自己与自己间不存在距离。此外主对角线两侧的元素是镜相对称的，即 $d_{ik} = d_{ki}$。距离系数矩阵是聚类分析的基础数据，聚类分析的过程是从距离系数矩阵出发的。

14.3.2　内积系数

前小节介绍用距离系数作为实体间的相似系数，原则上距离系数也可以用来表征变量间的相似程度，仅需要在 n 维的实体空间中计算变量间的距离。但实际上为表征变量间的相似程度用得更多的是各类内积系数，即夹角余弦，方差-协方差和相关系数等。

1. 夹角余弦。

原始数据矩阵(14-1)的每一列是一个变量在 n 维实体空间中的取值，也可以看成一

个 n 维的矢量。两个变量矢量的点积也称为它们的内积。

$$Q_{jk} = \sum_{i=1}^{n} x_{ij} \cdot x_{ik} \tag{14-12}$$

式子(14-12)给出变量 j 和变量 k 之间的内积系数。从基础三角知识可知,两个矢量内积的数值等于两个矢量的长度和它们间夹角余弦的乘积。即有

$$\cos\theta_{jk} = \frac{Q_{jk}}{\sqrt{Q_{jj}Q_{kk}}} \tag{14-13}$$

式(14-12)分母中的 Q_{jj} 和 Q_{kk} 分别是矢量 j 和 k 的长度。$\cos\theta_{jk}$ 是两个变量矢量间夹角的余弦,可以作为两个变量间的相似系数,当代表两个变量的矢量相重叠时,$\cos\theta_{jk} = 1$,而当两个矢量垂直时,$\cos\theta_{jk} = 0$。

2. 方差—协方差。

如果原始数据矩阵已经按公式(14-2)中心化,那么内积系数 Q_{jk} 是变量 j 和 k 的协方差 $\text{cov}(j,k) = ss_{jk}$ 的 $(n-1)$ 倍。因为协方差的定义是

$$ss_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{14-14}$$

也写作

$$\text{cov}(j,k) = ss_{jk} \tag{14-15}$$

而当 $j = k$ 时,公式(14-14)就是变量 j 和 k 的方差 $s_j^2 = ss_{jj}$ 和 $s_k^2 = ss_{kk}$。因此,量 $\frac{ss_{jk}}{\sqrt{ss_{jj} \cdot ss_{kk}}}$ 依然是两个变量矢量的夹角余弦,依然可作为变量间相似系数的度量。

3. 相关系数。

如果原始数据已经根据式(14-3)按标准差标准化了,即已转换为正规化的数据矩阵(14-5),那么方差 s_j^2 和 s_k^2 均等于1,协方差 $\text{cov}(j,k) = ss_{jk}$ 在数值上就等于两变量 j 和 k 间的相关系数了。因此变量间的相关系数是变量间相近程度的度量,也可以作为变量间的相似系数。

14.3.3 匹配系数和关联系数

上面两小节讨论的是数值变量的情况,分别介绍了用距离系数度量实体间的相似程度和用内积系数、协方差和相关系数等作为变量间相似程度的度量。本小节讨论二元变量的情况。考虑有2个实体,它们均被12个二元变量所描述,如表14-1所示。

表14-1 2个实体的12个二元变量的取值

	A	B	C	D	E	F	G	H	I	J	L	M
实体1	0	0	1	1	1	0	0	1	0	1	0	1
实体2	0	1	1	1	0	1	0	1	0	0	1	1

对表14-1的数据可以理解为2个动物群中12种物种的有无,也可以理解为2个墓葬中12种器物的是否存在。将表14-1的数据理解为12种物种在2个动物群出现情况的统计,并按交叉列联表形式整理如下:

表 14-2 12 种物种在 2 个动物群分布情况的交叉列联表

		实体 1	实体 1
		观测到的物种数	未见的物种数
实体 2	观测到的物种数	$a=4$	$b=3$
实体 2	未见的物种数	$c=2$	$d=3$

表 14-2 是我们已在第十章中见到的 2×2 交叉列联表。表中的 a 与 d 分别是在 2 个实体(动物群)中共同观测到的和都没有观测到的物种的数目,这些物种的发现与否对两个实体是共同的,它们表征两个动物群的共性。b 与 c 是仅在一个动物群中发现、而在另一动物群中未发现的物种数,表征两个动物群的相异性。因此可以用下列的各匹配系数来表征实体间相似的程度。

1. 简单匹配系数。

$$(a+d)/(a+b+c+d) \tag{14-16}$$

简单匹配系数的数值是在 0 与 1 之间波动,数值越大,表示两实体的性状间越接近。但是简单匹配系数有一定的缺点,因为某个物种不出现的情况可能比较复杂,例如在 8.5 节中曾论述到考古调查中没有发现某类物种不一定说明该类物种的确不存在。以比较早晚两期两座墓葬的相似程度为例,晚期的器物当然不能在早期的墓葬中出现,但也不是每种晚期器物都必须出现在每一座晚期的墓葬中。某种晚期器物在两座墓葬中的共同缺失并不一定反映这两座墓的共性。因此提出了另一种匹配系数,即 Jaccard 系数。

2. Jaccard 系数。

Jaccard 系数不考虑物种在两个实体中都不存在的情况,因此需要排除公式(14-16)中“d”。其表达式为

$$a/(a+b+c) \tag{14-17}$$

还可以定义有其他的匹配系数。它们都是在 0 与 1 之间变动。1 表示完全相似,而 0 为全不相似。

3. 各种关联系数和关联强度系数。

除了匹配系数外,还可以用二元变量间的各种关联系数来表征实体之间的相似程度。常用的系数有公式(10-3)定义的 χ^2 系数。

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(b+d)(a+c)(c+d)}$$

式中的 $n=a+b+c+d$。

还有公式(10-4)定义的 ϕ 系数

$$\phi = \sqrt{\frac{\chi^2}{n}} = \frac{|ad-bc|}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}$$

和公式(10-6)定义的 Yule's Q 系数

$$Q = \frac{ad-bc}{ad+bc}$$

14.4 等级聚类的原理、过程和问题

等级聚类有不同的聚类方法，或聚类策略。下面通过一个实例来说明等级聚类的过程，同时介绍聚类方法的选择。

表 14-3a 是 7 片原始瓷样品中 12 个元素含量的中子活化测量结果。其中 3 片出自江西吴城，3 片安徽苍圆塝和 1 片广东博罗梅花墩，都属于商周时期。这些数据是从第十五章表 15-28(86 片原始瓷中 19 个元素的含量值)中摘录的。

表 14-3a 7 片原始瓷样品中 12 种元素含量的中子活化测量结果

(K,Na 的测量单位为 %，其他元素为 μg/g)

	Ce	Cr	Eu	Hf	K	La	Lu	Na	Nd	Sm	Tb	Yb
吴 19	77.9	90.3	1.75	9.08	1.34	46.9	0.57	0.22	31.22	10.37	0.49	3.44
吴 20	81.1	85.2	1.8	9.27	1.39	44.5	0.59	0.25	33.07	11.43	0.58	3.5
吴 22	101.3	95.7	1.93	6.46	1.69	59.5	0.61	0.47	41.22	11.64	0.54	3.8
苍 57	93.7	39.5	1.3	8.56	2.12	50.4	0.39	1.41	51.4	6.01	0.77	2.21
苍 58	95.3	37.5	1.3	9.74	2.34	48.3	0.47	1.53	39.2	6.12	0.78	2.77
苍 59	89.9	33.1	1.25	3.26	2.05	49.8	0.36	1.42	38.4	5.98	0.79	1.95
博 07	141	74.4	1.81	10.6	2.17	73	0.54	0.10	66.2	10.1	1.44	4.93
平均值	97.17	65.10	1.59	8.14	1.87	53.22	0.50	0.77	42.96	8.81	0.77	3.23
标准差	20.96	27.40	0.29	2.50	0.40	9.90	0.10	0.64	12.14	2.65	0.32	1.02

对原始数据使用变量的标准差标准化，得各个 Z 分量值，数据列于表 14-3b。

表 14-3b 7 片原始瓷样品中 12 个元素含量用标准差标准化后的 Z 分量数据

	Ce	Cr	Eu	Hf	K	La	Lu	Na	Nd	Sm	Tb	Yb
吴 19	-0.920	0.920	0.532	0.376	-1.329	-0.633	0.655	-0.848	-0.967	0.591	-0.874	0.208
吴 20	-0.768	0.734	0.704	0.452	-1.216	-0.872	0.854	-0.809	-0.814	0.990	-0.598	0.267
吴 22	0.199	1.117	1.155	-0.671	-0.460	0.635	1.087	-0.464	-0.144	1.070	-0.719	0.561
苍 57	-0.165	-0.934	-0.992	0.169	0.626	-0.285	-1.157	1.001	0.695	-1.057	-0.008	-1.000
苍 58	-0.089	-1.008	-0.992	0.640	1.177	-0.497	-0.311	1.141	-0.310	-1.015	0.050	-0.450
苍 59	-0.347	-1.168	-1.163	-1.949	0.450	-0.345	-1.456	1.016	-0.375	-1.068	0.063	-1.256
博 07	2.090	0.340	0.755	0.984	0.751	1.997	0.327	-1.037	1.914	0.489	2.087	1.671

表 14-3b 是标准化后的原始数据。下面我们选择欧氏距离作为瓷片间的相似系数。7 片原始瓷样品两两间欧氏距离的计算结果列入表 14-4。

欧氏距离相似系数矩阵实际上是一个相异系数矩阵，主对角线上的元素都是 0，因为在化学元素组成的变量空间中每块瓷片自己与自己间是不存在距离的。主对角线两侧对称位置单元格中的数值是相等的，因为空间中两个点交换位置不会改变它们之间的距离。因此聚类过程中只需考虑主对角线右上方的元素，把相似系数矩阵作为三角阵对待。

表 14-4 7片原始瓷样品根据其元素含量的Z值计算欧氏距离的相似系数矩阵

		1 吴 19	2 吴 20	3 吴 22	4 苍 57	5 苍 58	6 苍 59	7 梅花 7
1	吴 19	0.000	0.683	2.555	4.970	4.720	5.502	6.358
2	吴 20	0.683	0.000	2.473	5.057	4.758	5.652	6.157
3	吴 22	2.555	2.473	0.000	5.198	5.040	5.613	4.998
4	苍 57	4.970	5.057	5.198	0.000	1.623	2.438	6.116
5	苍 58	4.720	4.758	5.040	1.623	0.000	3.059	6.094
6	苍 59	5.502	5.652	5.613	2.438	3.059	0.000	7.334
7	博 07	6.358	6.157	4.998	6.116	6.094	7.334	0.000

聚类的第一步是把两个最相近的实体聚成一组，从表 14-4 中看到 1 号和 2 号实体间的距离最短，是 0.683。因此第一步是把它们聚成一组，当成一个实体。第二步是要确定这个合并实体与其他实体之间的距离，例如怎样确定(1,2)合并组与 3 号实体之间的距离，这里有多种方法可供选择，称为不同的等级聚类方法。

14.4.1 等级聚类方法

本小节将通过计算(1,2)合并组与 3 号实体之间的距离，介绍几种常用的等级聚类方法。

1. 最近邻体法(Nearest neighbor)。由表 14-4 可见，实体对(1,3)和(2,3)的距离分别为 2.555 和 2.473。最近邻体法是从中选一个短的距离，即 2.473 作为合并组(1,2)与 3 号实体间的距离。最近邻体法又称简单联系(single linkage)法，它更广泛的意义是在两组实体间，选择一对最接近的实体，以这两个实体之间的距离作为这两个实体组之间相似程度的度量。

2. 最远邻体法(Furthest neighbor)。最远邻体法从(1,3)和(2,3)的距离 2.555 和 2.473 中选一个长的距离，即以 2.555 作为合并组(1,2)与 3 号实体间的距离。最远邻体法又称完全联系(Complete linkage)法，它与最近邻体法相反，在两组实体间选择一对差异最大的实体，以这两个实体之间的距离作为这两个实体组之间的距离。

显然，这两种方法有共同的缺点，在 1,2 号两个实体分别与 3 号实体的距离中只选择了其中的一个距离。在后面的聚类过程中只有被选的那个距离值将得到考虑，因此这两种方法分别使距离拉近变量空间被压缩或者把距离推远空间被扩张。

3. 组平均法(Between group average)。组平均法以(1,3)和(2,3)的距离 2.555 和 2.473 的平均值 2.514 作为合并组(1,2)与 3 号实体间的距离。因此组平均法又称为平均联系(average linkage)法。组平均法的一般计算公式是：

$$d_{c,a+b} = \frac{n_a}{n_a + n_b} d_{c,a} + \frac{n_b}{n_a + n_b} d_{c,b} \tag{14-18}$$

式中的 d 代表距离，n_a 和 n_b 分别是(a,b)实体组中实体 a 和 b 的数目。组平均法的计算工作量比上面 2 种方法大些，但克服了它们的上述缺点。

上面三种聚类方法都是适用于以欧氏距离作为相似系数的情况。下面介绍几种适

用于以欧氏距离的平方值作为相似系数情况的聚类方法。

4. 中线法(Median),它是在变量空间中代表 c 号实体的点与连接 a 和 b 实体点线段的中间点 e 之间距离的平方值作为合并组(a,b)与 c 号实体间的相似系数。具体的计算公式是

$$d^2_{c,a+b} = \frac{1}{2}d^2_{c,a} + \frac{1}{2}d^2_{b,c} - \frac{1}{4}d^2_{a,b} \tag{14-19}$$

式中的 d^2_{ij} 是 i,j 点间距离的平方。在简化的二维空间条件下,式(14-19)中各距离值的含义如图 14-1 所示,初等几何就能证明公式(14-19)的成立。

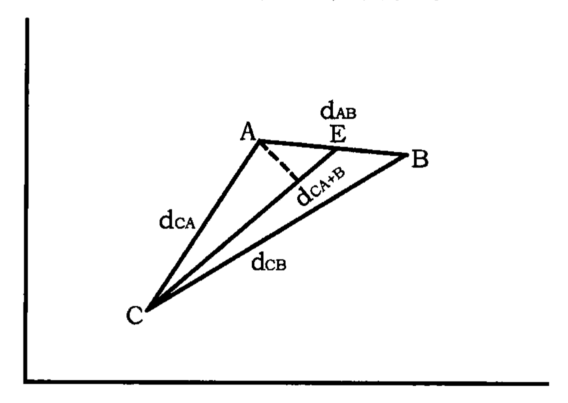


图 14-1 中线聚类法的示意图

5. 重心法(Centroid)。重心法是中线法的一种变型,它计算变量空间中代表 c 号实体的点与连接 a,b 实体点的线段的重心之间距离的平方值,并作为 c 号实体与合并组(a,b)间的相似系数。与中线法不同之处在于:不再将 ab 线段的中间点、而是 a 和 b 号实体的重心作为(a,b)合并组的代表。重心位置的确定需要考虑(a,b)组内各自包含的实体的数目。重心法的计算公式比式(14-19)略为复杂,这里不予写出。

6. Ward's 方法,又称平方和增量方法。它是基于方差分析的思想,使得每次合并后总的平方和的增加尽量小。其计算公式是

$$D_{c,a+b} = \frac{n_c + n_a}{n_c + n_{a+b}}D_{c,a} + \frac{n_c + n_b}{n_c + n_{a+b}}D_{c,b} - \frac{n_c}{n_c + n_{a+b}}D_{a,b} \tag{14-20}$$

式中 D 为相似系数(即欧氏距离的平方值),诸 n 值代表各相应组所包含的实体数。

在上述诸聚类方法中,Ward's 方法和组平均方法得到最广泛的应用,后者也称均值聚类方法。

14.4.2 等级聚类过程

聚类过程的出发点是由原始数据标准化后计算得到的实体间的相似系数矩阵。下面用组平均方法对表 14-4 的相似系数矩阵执行聚类过程。

(1) 前面已看到,第一步是将两个最接近的实体 1 和 2 聚合,聚合水平是 0.683。再用式(14-18)计算其他各实体与(1,2)合并组的相似系数,得到表 14-5a。

表 14-5a　对表 14-4 的数据第一步聚类的结果

	1,2	3	4	5	6	7
1,2	0.000.	2.514	5.014	4.739	5.577	6.258
3		0.000	5.198	5.040	5.613	4.998
4			0.000	1.623	2.438	6.116
5				0.000	3.059	6.094
6					0.000	7.334
7						0.000

表 14-5a 与表 14-4 比较，行与列的数目均少 1，而且第一行的数据发生了变化。

(2) 在表 14-5a 中，最小的数为 1.623，即最为相互接近的是实体 4 和实体 5。第二步是将 4 和 5 两实体聚合，聚合水平是 1.623，并根据公式 14-18 计算(4,5)合并组与其他各实体之间的相似系数，得到表 14-5b。

表 14-5b　第二步聚类结果

	1,2	3	4,5	6	7
1,2	0.000	2.514	4.877	5.577	6.258
3		0.000	5.119	5.613	4.998
4,5			0.000	2.749	6.105
6				0.000	7.334
7					0.000

(3) 聚类过程的第三步应该是实体 3 与(1,2)实体组聚合，聚合水平是 2.514。计算其他各实体与(1,2,3)合并组的相似系数，得到表 14-5c。这里需要注意的是：计算实体组(4,5)与实体组(1,2,3)的相似系数时，应考虑实体 3 是一个单独的实体，而实体组(1,2)包含 2 个实体。因此这个相似系数 $=\frac{2\times4.877+5.119}{3}=4.958$。同理，实体 6 与实体组(1,2,3)的相似系数 $=\frac{2\times5.577+5.613}{3}=5.589$。实体 7 与实体组(1,2,3)的相似系数 $=\frac{2\times6.258+4.998}{3}=5.838$。

表 14-5c　第三步聚类结果

	1,2,3	4,5	6	7
1,2,3	0.000	4.958	5.589	5.838
4,5		0.000	2.749	6.105
6			0.000	7.334
7				0.000

(4) 第四步应该是实体 6 与(4,5)实体组聚合，聚合水平是 2.749。计算其他实体与(4,5,6)组的相似系数，得到表 14-5d。计算相似系数时同样要考虑各组包含实体数的不同。实体组(1,2,3)与实体组(4,5,6)的相似系数 $=\frac{2\times4.958+5.589}{3}=5.168$，实体组(4,

5,6)与实体 7 的相似系数 $= \frac{2\times 6.105 + 7.334}{3} = 6.515$

表 14-5d 第四步聚类结果

	1,2,3	4,5,6	7
1,2,3	0.000	5.168	5.838
4,5,6		0.000	6.515
7			0.000

(5) 第五步是将(1,2,3)和(4,5,6)组合并,聚合水平为 5.168,得到表 14-5e。

表 14-5e 第五步聚类结果

	1,2,3,4,5,6	7
1,2,3,4,5,6	0.000	6.177
7		0.000

(6) 最后在 6.170 水平上,所有的实体聚合成一组。现将上面组平均方法的聚类过程总结在表 14-6 中。

表 14-6 聚类过程总结

合并步骤	组 1	组 2	聚合水平	聚类后的分组数
1	1	2	0.683	6
2	4	5	1.623	5
3	1,2	3	2.514	4
4	4,5	6	2.748	3
5	1,2,3	4,5,6	5.168	2
6	1,2,3,4,5,6	7	6.176	1

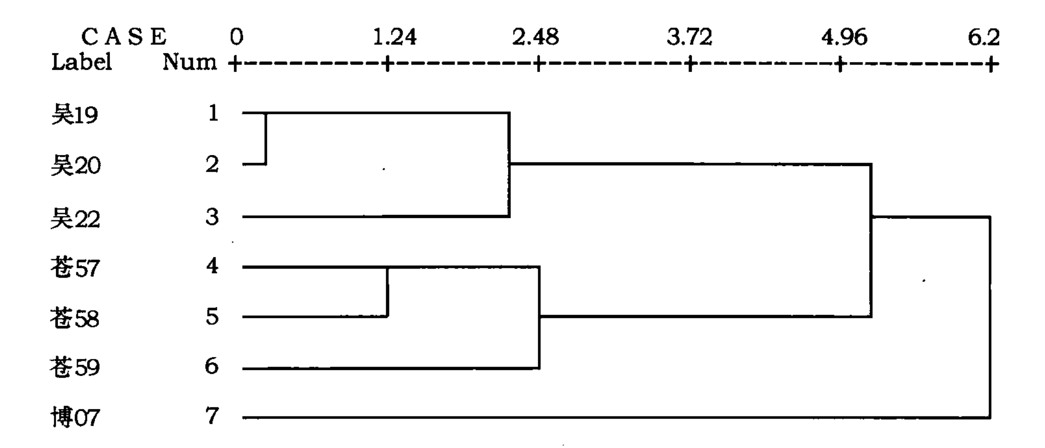


图 14-2 均值聚类方法对 7 片原始瓷片按其化学组成聚类的树枝状聚类图

前面聚类过程的最后结果用树枝状图 14-2 表示(因排版原因,该图需旋转 90 度观察)。底线上列出 7 个参加聚类的实体,不同高度处的水平线段反映在该聚合水平下,哪

两个实体或实体组聚合成一组。这样的水平线段共有6条,代表6次聚合过程。根据树枝状图最终应将这7个实体分成几类呢,这取决于分类界限值,或称聚类水平值的选择。如果分类界限值取在聚合水平2.748和5.168之间,那么7个实体被分成3类,分别是实体组(1,2,3)、(4,5,6)和 实体7。3类实体正好与瓷片的3个产地符合。如果将分类界限值提高到5.168和6.176之间,那么7片瓷片将分成2类,1—6号瓷片,即吴城和苍圆塝的6片瓷片共聚成一类,而博罗的1片瓷片自成一类。显然不应把分类的聚合水平选取得太低,使得分类的组数过多,分类组数过多显然是没有意义的。在这个演示例子中,瓷片的先验分类,即它们的产地是已知的,7片瓷片出土于三个地点。聚类分析的目的只是检验瓷片按其化学组成分类的结果与先验的产地知识是否一致。如果事先不掌握先验分类的知识,怎样确定分类界限值,即应该将全部实体分成几类,是需要认真考虑的问题,我们将在下一小节中讨论。

14.4.3　关于等级聚类的一些问题

等级聚类由于其数学原理简单明了,不需要对于原始数据的分布作任何前提假设,分类结果直观等特点,在各个学科都得到广泛的应用。但是在实际应用中也有一些问题是需要注意的。

1. 在一些实际问题中,我们事先并没有关于实体根据其外在性状分类的先验知识,那么在得到树枝状图后,应该怎样选取决定分类组数的聚合水平,应该将实体分成几类呢?这涉及类的概念和分类的标准,这方面并没有被普遍接受的共识。大致可以考虑以下几点:(1)每类的实体数不应太多;(2)每个类的个体间不应相互间差别太大,而各类的重心间应该有较大的差异;(3)分类的结果不应与常识相悖;(4)不同方法的分类结果不应差别太大。

2. 上面的第4点是十分重要的,聚类分析的结果依赖于相似系数和聚类方法的选择。选取不同的相似系数和使用不同的聚类方法可能会给出不同的树枝状聚类图,给出不完全相同的分类结果。图14-3和图14-4分别是用最近邻体法和Ward's方法对前述7片原始瓷的树枝状聚类图。与图14-2比较,可以看到,如果把7片瓷片分成3类,3种分类结果是一致的,正好与瓷片的3个产地相对应。但是如果把它们分成两类,Ward方法将吴城和博罗的瓷片合并为一类,苍圆塝为另一类,而其他两种方法却把吴城和苍圆塝归为一类,博罗的1片瓷片自成一类。

从3种聚类分析的结果看,如果将7片瓷片分成3类,聚类结果是稳定的,没有因聚类方法不同而改变,而且与外在的先验知识,即关于原始瓷片产地的知识也一致。至于在3类瓷片之间哪两类瓷片的化学组成更接近,3种聚类分析方法给出的结果是矛盾的。为了判别哪种聚类方案更符合实际,需要有更多的瓷片样品参加分析,并与其他的分类方法的结果进行比较。

在这个例子中,实体数目很少,仅7片瓷片,而且各地瓷片的化学组成差别也较大,因此聚类分析的结果比较稳定。当参加分析的实体数目很大时,不同的聚类方法一般不可能给出完全一致的树枝状聚类图,仅能希望的是得到“大同小异”的分类结果。为了希望分类结果能真实反映数据的内在结构,而不是分类方法的误导,作者强烈提倡同时用

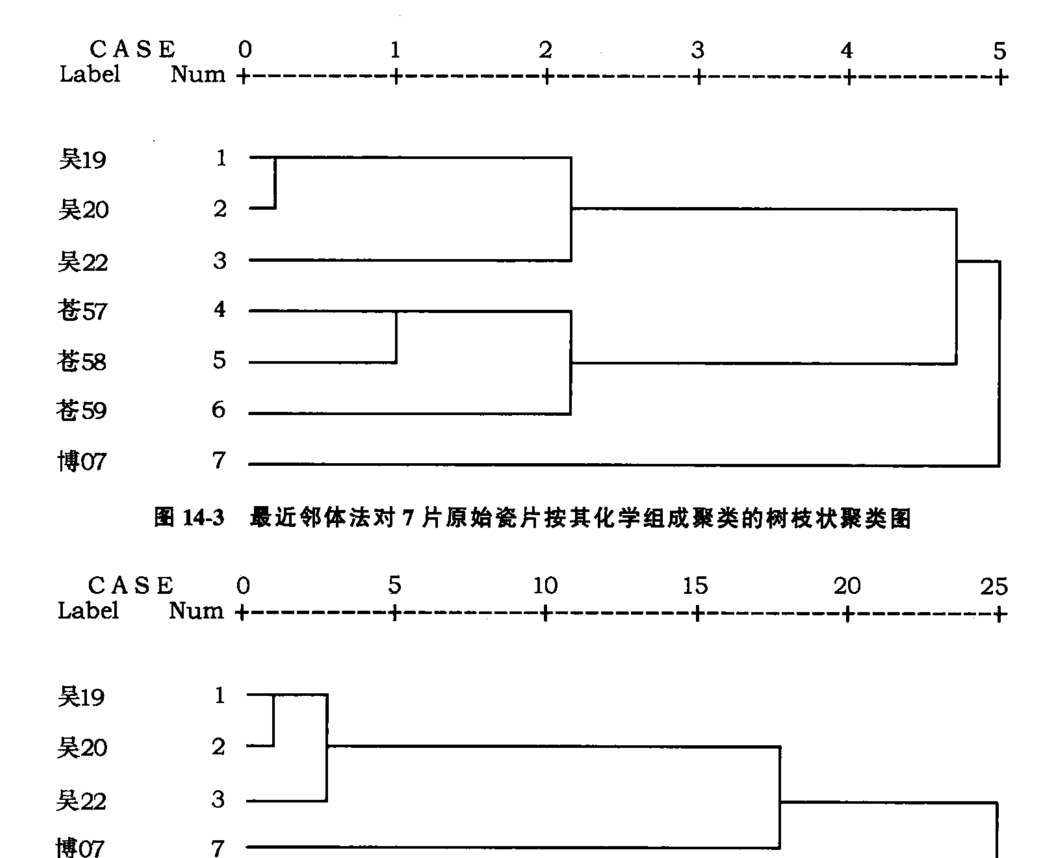


图 14-3 最近邻体法对 7 片原始瓷片按其化学组成聚类的树枝状聚类图

图 14-4 Ward's 方法对 7 片原始瓷片按其化学组成聚类的树枝状聚类图
（聚类水平未按比例显示）

两种或两种以上的方法对同一组数据进行分类，包括等级聚类、非等级聚类、分划以及主成分分析等多种分类方法。以便能有较高的置信度确认分类结果的稳定性和可解释性。

单种方法等级聚类的结果有时会误导的。Wright(1989)曾在平面上随机产生了一些二维的点（x_i，y_i），如图 14-5a 所示，这些点基本上是连续、均匀地分布的。但是用 Ward 的方法对它们进行聚类，却明显地分成二类(见图 14-5b)。

在这个例子中，聚类方法对本身并没有分组结构的数据给出了实际不存在的分组结果；另一方面聚类分析并不能保证能完全正确地揭示数据中确实存在的分组结构。例如，在后面 14.5 节对殷墟颅骨的分类研究中，虽然均值聚类正确地对 22 组颅骨分成北亚，东亚和高加索等 3 组，但是 Ward 方法却未能将东亚和高加索两组颅骨分开。鉴于聚类方法还有下面将介绍的其他一些缺点，尽管不断有使用聚类方法分析考古数据，特别是科技考古数据的文章发表，有部分西方考古学家怀疑聚类分析方法处理考古资料的能力。两本英国出版的关于多元统计分析应用于考古研究的专著：Baxter(1994)书的第八章

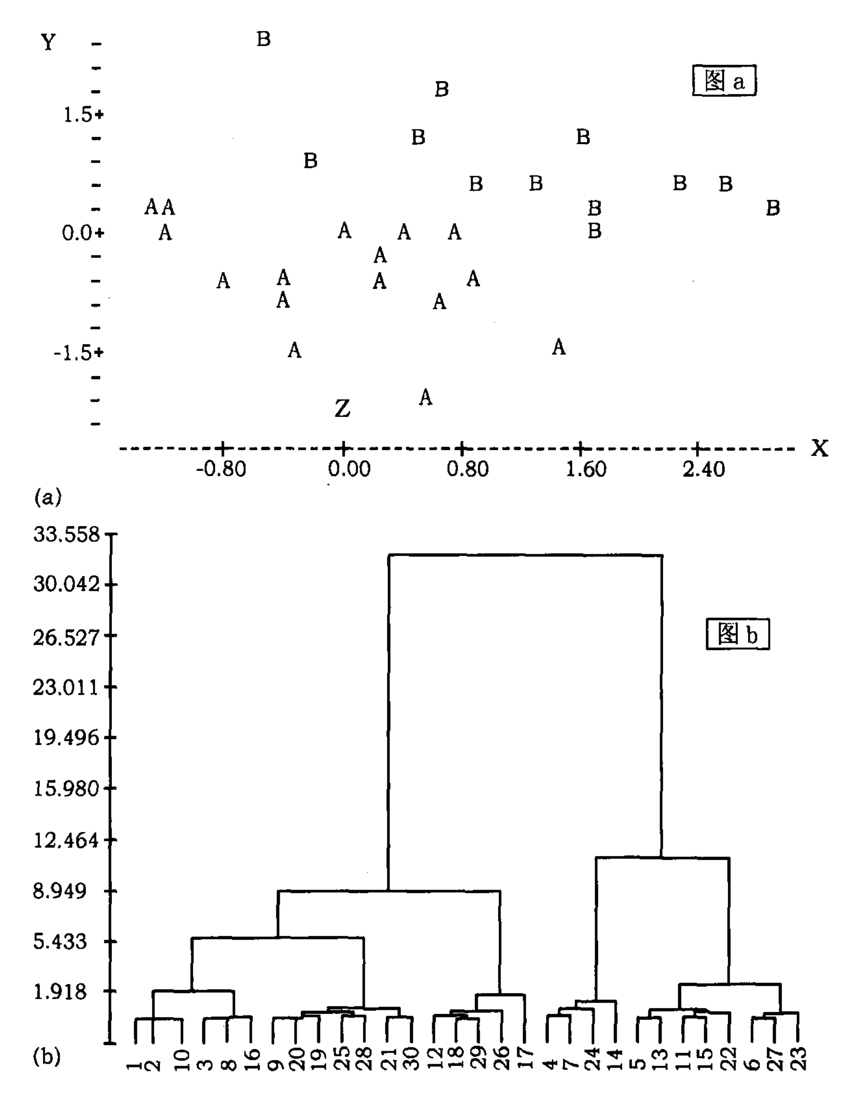


图 14-5　Ward 方法对随机产生的点进行聚类的结果：图 a 是随机点的散点图，图中的点分别用聚类分析的分类类别 A 和 B 标记。图 b 是聚类树枝状图，明显分为 A 和 B 两类。

和 Shennan(1997)书的第十一章，都详细地讨论了聚类方法的优缺点，他们强调要审慎地对待由单一聚类方法给出的分类结果，建议聚类方法与其他分类方法的结合。这点是十分重要的。

3. 等级聚类还有其他的弱点值得我们注意。例如，如果实体 a 和 b 在聚类的初期阶段被分在同一组，那么在整个分类过程中 a 和 b 将一直在一起，不能进行调整。如果某个实体因为某种偶然因素被错误地分到某一组，那么这个错误会自始至终地影响后面的聚类过程，有时会导致严重的后果。在下一章中将介绍 K－means 方法分类中，实体在初始的分类组之间是可以调整的。K 方法给出的是网状的分类图，而不是树枝状的。

等级聚类的另一缺点是不易看出原始数据中的各变量对树枝状聚类图的建立的作用，即难以分析每个变量的取值范围和分布在分类中的作用。而在主成分分析中通过因

子负载,在判别分析中通过判别函数能清楚地,而且定量地揭示每个变量在实体分类中的贡献。因此从这个角度我们也提倡在对考古资料和科技考古资料数据进行等级聚类分析时,尽量同时使用主成分分析等其他多元统计方法,以便能诠释原始数据中各变量与分类方案间存在的因果关系。

4. 等级聚类的优点是,它不仅对实体进行了分类,而且因为每个实体的聚类过程有先后次序,从而产生的树枝状的聚类结构能给出补充的信息。如果被分类的实体是生物学的物种,那么树枝状的系统结构能提供关于生物种属的发生和演化的信息。但是聚类分析应用于古陶瓷的产地溯源中,树枝状的系统结构的含义并不容易解释,各地陶瓷的化学组成间不存在演化的关系。

14.5 等级聚类应用实例:安阳殷墟颅骨的种系分类研究

1928—1935 年在安阳殷墟进行了举世瞩目的 12 次考古发掘,其成果之一是在西北岗祭祀坑中出土了 300 多具古人的颅骨,这批材料于 40 年代末被运往台湾。杨希枚(1985)对这些颅骨进行了研究,认为它们的某些测量项目和颅骨指数变异甚大,超出了同种系人种同类项目的标准差。鉴于祭祀坑颅骨的成分可能比较复杂,有战争的俘虏,有虏获的奴隶,杨希枚提出了这批颅骨代表异种系群体的观点,并根据颅骨的形态进一步把它们分成 5 组。第一组代表典型的蒙古人种北亚类型。第二组与现代生活在大洋洲的美拉尼西亚人和巴布亚人相近。第三组只有 2 个个体,属高加索人种。第四组代表蒙古人种北亚极区的爱斯基摩类型。第五组的种系类型尚待研究。但是韩康信等提出了不同的意见,他们虽同意杨希枚所划分的第一组颅骨接近蒙古人种北亚类型,但认为第二组属蒙古人种南亚类型,而其他三组之间不存在显著的差别,均为蒙古人种东亚类型。因此韩康信等(1985)反对异种系的观点,认为殷墟祭祀坑出土的颅骨总体上仍属纯种系,均属于蒙古大人种系。韩康信等还对解放后发掘的殷墟中小墓出土的颅骨进行了测量研究。殷墟中小墓的主人应属殷商的自由民,应代表殷民族颅骨的特征。韩等的研究表明,中小墓颅骨间的差异不大,应属蒙古人种的东亚类型。同时他们也注意到中小墓颅骨中有 8 个颅骨具有颧骨间偏宽和颅高偏低等北亚类型的特征,韩等用“殷中小墓 II”来标识这 8 具带有北亚类型形态的颅骨。

为了鉴定这些颅骨间,包括祭祀坑颅骨和殷墟中小墓颅骨的种族关系,为了判断殷墟祭祀坑颅骨是异种系的还是纯种系,需要对中小墓颅骨与祭祀坑颅骨的之间的各项测量指标,对它们与我国中原以及北亚地区各时段的颅骨的各项测量指标进行对比研究。表 14-7 列出了 22 组人群的 21 项颅骨测量指标的平均值。

人类学的传统方法是通过很多个单项指标间的比较来研究人群间的种系关系,而本书的作者(1991)曾尝试利用多元分析方法中的聚类分析和主成分分析对表 14-7 的数据进行了综合的研究。聚类分析的结果显示在图 14-6 和图 14-7 两张树枝状聚类图中。前者用欧氏距离作为相似系数和均值聚类方法,后者用欧氏距离平方作为相似系数和 Ward 聚类方法,两个聚类过程都是首先对表 14-7 的原始数据按标准差进行了标准化。聚类是使用 SPSS 软件完成的。

表 14-7　22 组颅骨(男)测量数据平均值(根据韩康信等(1985))

	颅长	颅宽	颅高	耳上颅高	最小额宽	颧宽	上面高	鼻高	鼻宽	眶宽	眶高	面角	齿槽点角	鼻根点角	颅指数	颅长高指数	颅宽高指数	中上面指数	鼻指数	眶指数	额宽指数
殷祭祀坑 I	182.5	144.4	135.1	115.4	93.8	141.2	73.6	54.5	27.3	41.6	33.5	85.8	69.8	65.9	79.2	74.2	93.6	70	50.4	80.8	64.8
殷祭祀坑 II	182.9	136.9	141.1	119.3	92.7	134.5	71.4	51.6	27.7	41	32.1	83.1	70.7	67.7	75.1	77.1	104.1	70.7	54.4	78.6	68
殷祭祀坑 III	181.5	133.5	138.5	116.3	92.3	131.5	71	52.8	25	40.5	32.5	84	72.8	66	73.6	76.3	103.8	70.7	74.4	80.2	69.1
殷祭祀坑 IV	182.5	139.1	140.1	118.4	91.2	135.1	72.9	52.8	26.9	41.9	33.1	86	71.9	65.2	76.4	76.5	101.3	71.8	50.7	79.1	65.8
殷祭祀坑 V	180.1	136.8	136.7	116.6	89	131.3	72.2	52.9	26.5	40.4	32.8	83.3	70.8	65.9	75.7	76.3	100.2	72.4	50.3	81.1	65.1
殷中小墓 I	184	140.1	140.3	117.8	90.4	133.1	73.8	53.4	27	42.4	33.6	83.8	72	67.1	76.5	76.1	99.4	70.6	51	78.6	64.4
殷中小墓 II	187.2	142.7	134.8	115.1	93.9	145.4	75.1	56.4	29	44.9	35.5	84.6	72.2	67.1	76.3	72.1	94.5	72.5	51.4	79.3	65.5
史前华北	181.6	137	136.8	116.4	92.3	130.7	74.8	55	25.6	45	33.8	85	71.7	64.7	75	75.7	100.5	73.9	47.3	75.1	67.4
柳　湾	185.9	136.4	139.4	118.3	90.3	137.2	78.2	55.8	27.3	43.9	34.3	89.2	72.2	65.4	73.9	74.7	101	73.4	49.1	78.5	65.9
仰韶(合并)	180.7	142.6	142.5	121.6	93.6	136.4	73.4	53.4	27.6	43.4	33.5	81.4	69.2	69.4	79.1	78.6	99.4	68.3	52.1	77.2	65.6
现代华北	178.5	138.2	137.2	115.5	89.4	132.7	75.3	55.3	25	44	35.5	83.4	69.5	64.9	77.6	77	99.5	77	45.3	80.7	64.7
现代华南	179.9	140.9	137.8	116.9	91.5	132.6	73.8	52.6	25.3	42.1	34.6	84.7	70.5	69.5	78.8	77	97.8	74.3	47.4	84.9	64.9
现代蒙古	182.2	149	131.4	114.4	94.3	141.8	78	56.5	27.4	43.2	35.8	87.5	69	64.6	82	72.1	88.2	75.9	48.6	82.9	63.3
爱斯基摩 东南 I	181.8	140.7	135	113.9	94.9	137.5	77.5	54.6	24.4	43.4	35.9	83.8	68.2	67	77.6	74.3	96	78.3	44.8	83	67.5
爱斯基摩 那俄康 II	183.9	143	137.1	116.7	98.1	140.9	78.2	55.7	23.5	44.5	35.9	85.6	67.6	67.5	77.5	74.6	95.9	77.8	43	80.8	68.6
爱斯基摩 近代 III	182.3	141.2	135.2	113.9	94.9	138.4	77.6	54.6	24.4	43.4	36.1	83.8	68.2	67	77.6	74.3	96	78.3	44.8	83	67.5
楚克奇滨河 I	182.9	142.3	133.8	112.1	95.7	140.8	78	54.7	24.6	44.1	36.3	83.2	68.8	66.2	77.9	73.2	94	76	44.7	82.4	67.3
楚克奇驯鹿 II	184.4	142.1	136.9	113.8	94.8	140.8	78.9	56.1	24.9	43.6	36.9	83.1	68.3	66.8	77.2	74.2	96.3	77.9	44.5	84.5	66.7
布列亚特 西 I	183.6	147.5	135.4	115.4	96.5	143	79.1	56.4	26.8	42.9	35.7	86.9	69.6	64.3	80.5	73.8	91.8	76	47.6	83.3	65.4
布列亚特 东 II	181.7	150.3	132.6	114.5	94.9	142.6	76.9	55.5	26.6	42.3	35.3	88	70.1	64.8	82.7	73	88.2	76.1	48.2	83.3	63.1
布列亚特 外贝加尔 III	181.9	154.6	131.9	115.5	95.6	143.5	77.2	56.1	27.3	42.2	36.2	87.7	70.6	64.2	85.1	72.5	85.3	75	48.7	83	61.8
布列亚特 高加索	180.4	140.9	131.3	114.3	96.2	128.3	70.6	51	24.1	45	33.5	85.6	71.7	66	78.2	72.5	93.1	77.1	47.5	74.8	68.3

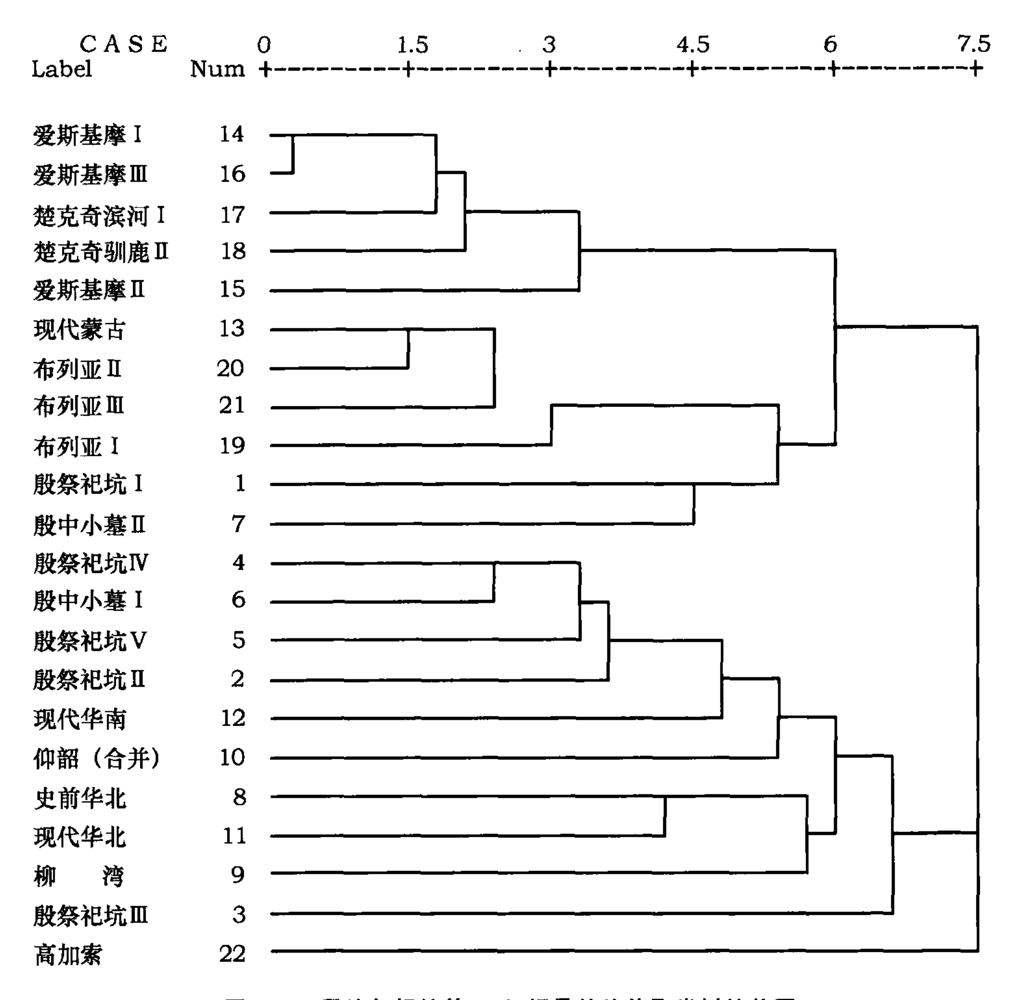


图 14-6 殷墟祭祀坑等 22 组颅骨的均值聚类树枝状图

在图 14-6 上,取聚合水平大于 6.5 时,22 组人群可分为 3 组。第一组包含爱斯基摩、楚克奇、布列亚和现代蒙古人等全部北亚类型的人群,但专门挑选的其颅骨形态接近北亚人群的殷墟祭祀坑 I 和中小墓 II 也进入此组。现代华北、现代华南、史前华北、古代的仰韶、柳湾、以及除殷墟除祭祀坑 I 和中小墓 II 外的其他殷墟颅骨都进入第二组,显然这一组代表几千年以来的东亚人群。第三组只包括高加索人一个组。如果将作为分类标准的聚合水平降低为 4.5 左右,那么北亚人群又可进一步分成:(1)北亚极地爱斯基摩和楚克奇人,(2)典型蒙古北亚类型的布列亚和现代蒙古人和(3)祭祀坑 I 和中小墓 II 等三小类。在第 2 组东亚人群中,除形态像高加索人的殷墟祭祀坑 III 颅骨外,其他 4 组殷墟颅骨均合在一起。因此由图 14-6 可清楚地看到,按 21 项颅骨测量数据均值聚类的结果与人类学关于欧亚人种种系的知识是一致的,并不支持殷墟祭祀坑颅骨异种系的观点。除殷墟祭祀坑 I 和殷墟中小墓 II 的颅骨下面将进一步讨论外,殷墟祭祀坑和中小墓的其他多数颅骨的形态一致,并均属于蒙古大人种东亚类型。

均值聚类的分析结论基本上得到 Ward 方法聚类分析的支持。Ward 方法聚类的图

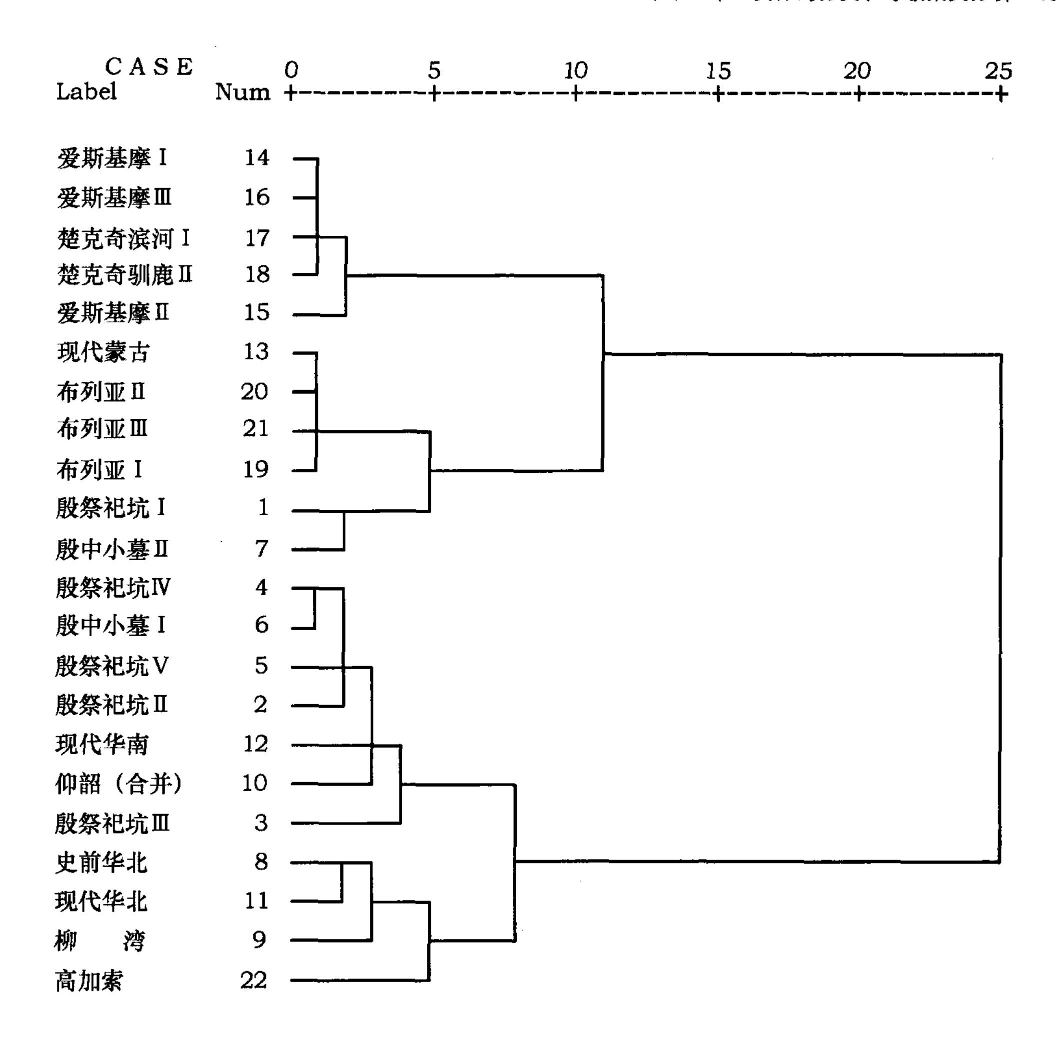


图 14-7 殷墟祭祀坑等 22 组颅骨的 Ward 方法聚类树枝状图

14-7将 22 组人群分成 2 组，与图 14-6 的唯一的不同处在于它将高加索组与东亚人群组合并为一组。但只要降低聚合水平，高加索组与东亚各组是可以分开的。

本书第十七章将介绍主成分分析方法，这里我们给出这 22 组颅骨测量数据（表 14-7）的主成分分析结果。它们用两张主成分散点图表示。在这两张散点图第一主成分能明确地分辨东亚和北亚人群，代表后者的样品点均处于图的右半部分，其第一主成分值均为正；而前者除殷墟祭祀坑Ⅰ和殷墟中小墓Ⅱ外均处在图的左半部分，其第一主成分值均为负值。对于从殷墟颅骨中专门挑选的其形态接近北亚类型的祭祀坑Ⅰ和中小墓Ⅱ颅骨，虽在这两张图中其样品点位置偏右，第一主成分值为正，但其绝对值很小，离代表真正的北亚人群的诸点有相当的距离。考虑到同种系人群中的个体差异和颅骨各项指标本身的涨落，殷墟祭祀坑Ⅰ和中小墓Ⅱ颅骨还是应属于东亚人群，它们与殷墟的全体颅骨是属于同一种系的。在图 14-9 可看到高加索组与其他 21 组人群在第三主成分上是有显著差别的。主成分分析能揭示原始数据的变量对各个主成分的贡献。高加索人颧宽和鼻宽狭以及颅高较低的特征导致其第三主成分值小，为负值。即综合考虑前三个主成

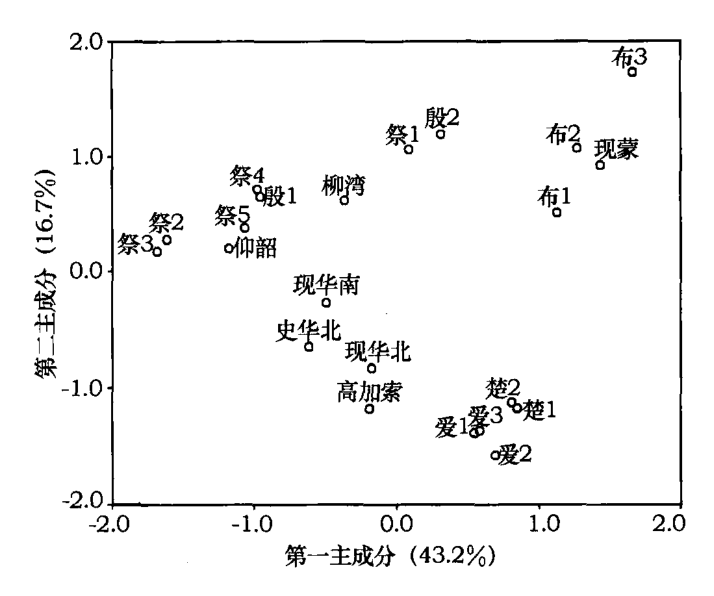


图 14-8　殷墟祭祀坑等 22 组颅骨的第一和第二主成分散点图

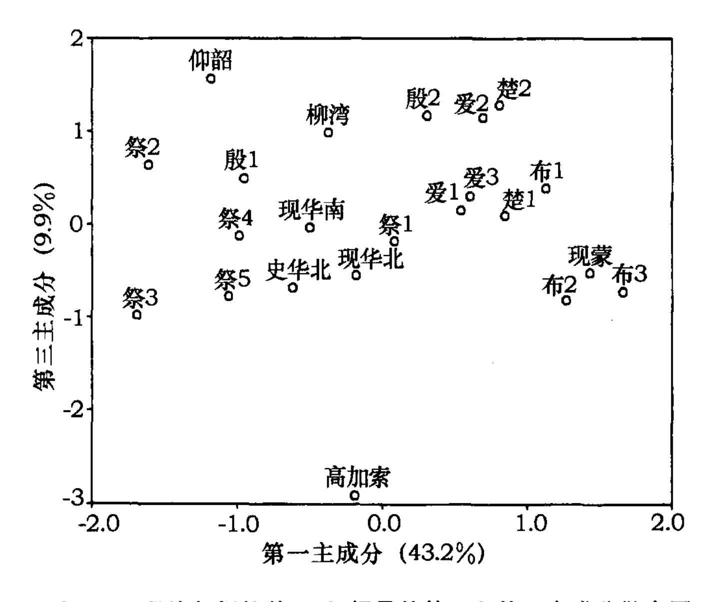


图 14-9　殷墟祭祀坑等 22 组颅骨的第一和第三主成分散点图

分，则东亚、北亚和高加索三种类型的人群能清楚地区分开，与均值聚类的结果一致。

总之，聚类分析和主成分分析两种多元方法对 22 组颅骨 21 项测量指标的研究结论是相互印证、相互补充的，并支持韩康信等关于殷墟祭祀坑颅骨属纯种系的观点。多元分析的研究结论比人类学研究中传统的通过多个单项指标间比较研究给出的结论更直观，更简明。因为后者是一系列单项研究结果的表述，而前者是研究结果的综合表述。同时这个研究实例也显示了同时使用两种或两种以上多元分析方法对实体正确分类的重要性。

14.6　单元等级分划

如果说聚类分析是将实体或实体群逐步两两聚合的综合过程,那么等级分划是将全部实体逐步分划的分解过程。等级分划是根据一定准则将全部 n 个实体划分成 2 组,然后再对其中的每一组再一分为二,这样重复进行,直到可以认为已分划的各子组内的实体已经是同质的,不需要再进一步分划。如果主要根据单个变量的取值作为划分的准则,称为单元的分划,当然也要考虑该变量与其他变量间的关系。本节仅讨论单元分划,而且变量属于二元变量的情况,下面将通过一个实例来讨论。

假设有 8 个墓葬,其中观察到有 A—F 等 6 种器物,表 14-8 是原始数据统计表,统计了这些器物在墓葬中的分布,“1”表示存在,“0”为缺失。任务是要根据器物的分布对 8 座墓葬进行分组。单元分划的第一步是要在 A—F 等 6 种器物中确定一种有分类意义的器物,根据它的存在与否将墓葬群分成 2 组。这种具有分类意义的变量在生物分类学中称为临界种。

表 14-8　A—F 等 6 种器物在 8 个墓葬中的分布,“1”表示被发现,“0”为未被发现

墓号	A	B	C	D	E	F
1	1	1	0	1	0	1
2	1	0	1	1	0	0
3	1	0	0	1	0	0
4	0	1	1	0	1	1
5	1	1	0	1	1	1
6	1	0	0	1	0	0
7	1	1	0	1	0	1
8	0	0	1	1	1	1

14.6.1　分类变量的确定

有多种方法或准则来确定分类变量,最常用的方法是利用关联系数和信息系数,后者要求分组后信息量的减少最大。这里不准备讨论怎样定义一组数据的信息量,以及分组和并组过程中信息量是怎样改变的等,即不讨论怎样用信息系数来确定分类变量。本节仅介绍怎样利用关联系数来确定分类变量。

在诸变量中,应该选择与其他的变量之间关联最强的变量作为分类变量,这样当以它的存在与否来分组时,其他变量也已经尽可能充分地被考虑了。可以利用第十章公式(10-3)定义的 $\chi^2 = \dfrac{n(ad - bc)^2}{(a+b)(b+d)(a+c)(c+d)}$ 值来检验变量之间是否关联,用公式(10-4)定义的 $\phi^2 = \dfrac{\chi^2}{n}$ 来度量变量之间的关联强度。为了确定分类变量,首先需要计算 6 个变量两两之间的 χ^2 值和 ϕ^2 值,各有 15 个。为此写出 8 个实体对于变量 A,B 分布的 2 ×2 列联表:

		B	B
	状态	存在	未见
A	存在	1	1
A	未见	3	3

可以计算得到 A,B 两变量间的 χ^2 值:$\chi^2_{AB} = \dfrac{8 \times (1 \times 3 - 1 \times 3)^2}{(1+1) \times (1+3) \times (1+3) \times (3+3)} = 0$。同样对于变量 A,C 有

		C	C
A	状态	存在	未见
A	存在	0	2
	未见	5	1

计算得到 $\chi^2_{AC} = \dfrac{8 \times (0 \times 1 - 2 \times 5)^2}{(0+2) \times (2+1) \times (0+5) \times (5+1)} = 4.444$ 。

用同样方法可以计算得到 $\chi^2_{AD} = 0.889, \chi^2_{AE} = 4.444, \chi^2_{AF} = 1.600 \cdots\cdots \chi^2_{EF} = 2.880$ 等共 15 个 χ^2 值。相应的 $\phi^2_{AB} = \dfrac{\chi^2_{AB}}{8} = 0, \phi^2_{AC} = \dfrac{\chi^2_{AC}}{8} = 0.566, \cdots\cdots \phi^2_{EF} = \dfrac{\chi^2_{EF}}{8} = 0.360$ 等共有 15 个 ϕ^2 值。第十章曾说明 ϕ^2 的取值范围是在 0 与 1 之间,这里再规定:每个变量自身的关联强度 ϕ^2 值为 $\phi^2_{JJ} = 1$。

下一步是在一定的显著性水平下检验变量之间的关联是否显著。因为是 2×2 的列联表,自由度为 1,对应于 0,15, 0.10 和 0.05 显著性水平的 χ^2 的临界值相应为 2.07, 2.706 和 3.841。如果选取 $\alpha = 0.15$,而变量 J 与 K 之间的 $\chi^2_{JK} < 2.07$,那么在 $\chi = 0.15$ 的显著性水平下变量 J 与 K 之间的关联没有统计意义,可以不必去考虑,即认为相应的 $\phi^2_{JK} = 0$。这样可以对所有的 ϕ^2_{JK} 值列出一个 6×6 的矩阵,我们用表格的形式列出如下:

	A	B	C	D	E	F
A	1	0	0.556	0	0.556	0
B	0	1	0	0.333	0	0.6
C	0.556	0	1	0	0	0
D	0	0.333	0	1	0	0
E	0.556	0	0	0	1	0.36
F	0	0	0	0	0.36	1
列和 Si	2.112	1.333	1.556	1.333	1.916	1.96

上面表格的最后一行是各列数值的和,是每个变量与所有 6 个变量(包括自身)的关联强度的总和 S_i。应该选择 S_i 最大的那个变量作为分类变量,因为在 6 个变量中它与其他的变量之间的关联最强。在本实例中第一列的列和最大,为 2.112,因此应该选第一个变量,即 A 型器物作为分类变量。

需要说明一种特殊情况,如果在表 14-8 中有某个变量,它对于 8 个实体的取值全是 1 或全是 0(某种器物在 8 座中全出现或全缺失),那么这个变量与别的变量间的关联系数

是无法计算的,可以规定它和其他变量间的关联系数为0。实际上这类变量在分类中是不起作用的,可以将其剔除。

14.6.2 分划过程

(一)第一次分划的结果。

对表14-8的8个实体按变量A取值为1和0划分,将分成2组,第一分组包含1,2,3,5,6,7等6个实体,它们的第一个变量,变量A的取值都是1;第二分组由实体4和8组成,都是A为0的实体。表14-9列出第一分组的数据。

表14-9 B-F等5种器物在A=1的6个墓葬中的分布

墓号	B	C	D	E	F
1	1	0	1	0	1
2	0	1	1	0	0
3	0	0	1	0	0
5	1	0	1	1	1
6	0	0	1	0	0
7	1	0	0	0	1

下面对第一分组中的6个实体进一步分组。需要从B-F等5个变量中再选择一个分类变量。同样方法先计算各变量之间的χ^2值,计算结果只有B与F间的$\chi^2_{BF} = 6 > 2.07$,其他的χ^2_{JK}值均小于2.07,相应的ϕ^2_{JK}值均应为0。这样5×5的ϕ^2_{JK}矩阵为:

	B	C	D	E	F
B	1	0	0	0	1
C	0	1	0	0	0
D	0	0	1	0	0
E	0	0	0	1	0
F	1	0	0	0	1
列和 Si	2	1	1	1	2

由这个矩阵数据可知,应选择变量B或F为分类变量,从表14-9可见这两个变量之间是完全的关联,选择其中的任意一个是等效的。

(二)第二次划分的结果与讨论。

按照变量B的取值,第一分组的6个实体又进一步分成2组,实体1,5,7为一组,它们的$B = F = 1$;而$B = F = 0$的实体2,3,6被划分为另一组。

经过2次分划,8个墓葬分成3个分组。对于每个分组需要继续计算变量间的χ^2值,分析是否需要进一步地划分。实际计算得到的χ^2值均小于2.07,说明不需要对这3个分组再作划分,每个分组中的实体可以认为是同质的了。变量间χ^2值的统计检验为分划过程提供了一个实际上的终止规则,当所有的χ^2值统计上不显著时,分划过程也就停止了。

整个分划过程的结果可以用树枝状图14-10来总结,图中从上而下记录了2次分划

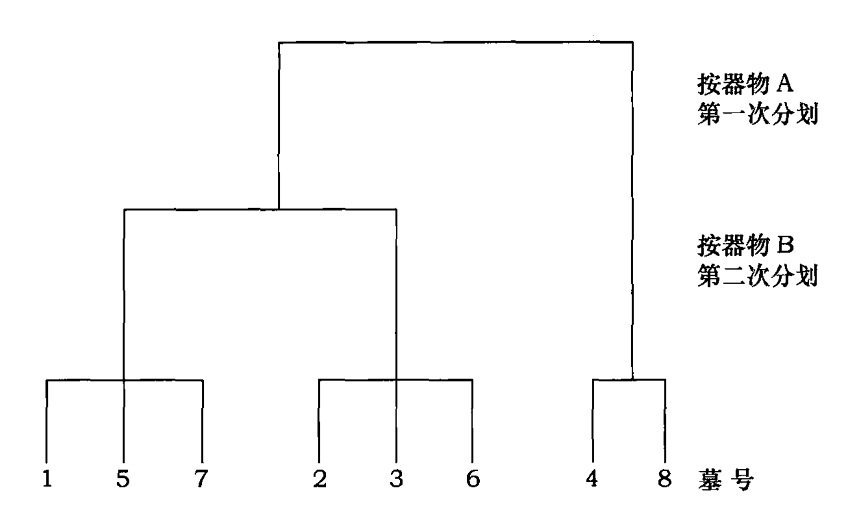


图 14-10　根据墓中器物的分布用关联系数对 8 个墓葬分划的树枝状图

所依据的变量(器物)名称。在底线上标明了每一组的实体编号。水平的分划线的高低应与每次分划的水平,即 ϕ^2 的列和值相当。

对原始数据表 14-8,按照分划结果对实体重新排列整理,得到表 14-8a。

表 14-8a　A-F 等 6 种器物在 8 个墓葬中的分布,墓葬已按分划结果分组排列

		A	B	C	D	E	F
第	1	1	1	0	1	0	1
一	5	1	1	0	1	1	1
组	7	1	1	0	0	0	1
第	2	1	0	1	1	0	0
二	3	1	0	0	1	0	0
组	6	1	0	0	1	0	0
第三	4	0	1	1	0	1	1
组	8	0	0	1	1	1	1

由表 14-8a 可见,第一组 1,5,7 三个墓葬对 4 种器物 A,B,C,F 是同质的,即这 4 个器物变量对该组的 3 个实体取值均为 1 或均为 0。第二组 2,3,6 三个墓葬对于总共 6 种器物中的 5 种是同质的,仅对于器物 C 不同质,同质的比例高达 84%。第三组 2 个墓葬对 4 种器物同质。高的同质比例说明上面进行的分划是有效的。

上述墓葬群根据是否含有某种分类意义器物作单元分划,与传统的考古研究中根据典型器物对墓葬的分类或分期相比较,它们在思维逻辑上颇为相似。分析计算一个墓葬群中器物间的关联强度与传统的考古研究中寻找墓葬中较为固定的器物组合,它们在研究目标上也很接近。因此,单元划分方法在考古研究的墓葬分期中应该有应用的前景。

单元分划方法的应用也受到一定程度的质疑,主要的问题是应该怎样对待(0,0)匹配。这在 14.3 节介绍 Jaccard 匹配系数时已提及。国外的一些文献中比较提倡用信息系数来确定单元分划中的分类变量,但信息系数的计算工作量可能稍大。

总之,无论是等级聚类或等级分划在考古实体的分类研究中都有应用前景,但同时

其本身也有一系列问题需要进一步探索。

14.7 非等级的 K 均值分类方法

14.7.1 K 均值分类方法的原理和执行过程

K 均值分类方法是一种非等级的分类方法,它的英文名称是 K-means 分类方法。假设有 n 个实体,每个实体被 m 个变量所描述,如表 14-1 所示,需要用 K 均值方法对实体进行分类。K 均值分类方法首先需要对原始数据作标准化处理,具体分类过程大致有下列步骤:

(1) 首先要确定对实体群分成几组,这要依据对实体群全部测量数据的已有的知识来确定。当然如果第一次的分类结果被认为不合适,可以重新规定分组数目再行分类。

(2) 确定初始类中心。在确定分类的数目后(譬如说分成 k 类),还要指定各个类中心的坐标值,称为初始类中心。k 个初始类中心可以由研究人员给定,也可以由计算机自动生成。计算机一般选择 k 个实体的坐标值作为类中心,并在选择时考虑它们之间的距离应适当拉开。

(3) 归类或分派。无论是每个实体,或者是类中心都是 m 维空间中的一个点,从而可以计算点与点之间的距离,一般用数据标准化后的欧氏距离。归类或分派是将实体一一归到与自己距离最短的类中心所在的类中,完成了第一次分类。归类程序也有两种,一是归类过程不改变原来的类中心的位置,二是当一个实体归到某类后重新计算该类中心的位置作为该类实体新的中心。后一种方法称为使用 running means 归类,计算量相应会大些。

(4) 迭代和迭代终止规则。完成第一次的归类后重新计算各类的中心值,然后再将每个实体一一归到与自己距离最短的新的类中心所在的类中,完成了一次迭代过程。反复地迭代计算,直到完成事先规定的迭代次数或者满足规定的迭代收敛标准。所谓收敛标准可以这样规定,譬如要求二次迭代前后 k 个类中心距离改变的最大值不大于初始类中心间最短距离的百分之一。当然,如果前后两次迭代计算的结果不改变实体的分类,迭代过程也就自然停止了。迭代过程收敛的快慢与分类数目的设定,与初始类中心位置的选择是否合适有关。

(5) 分类结果的分析。在分类过程结束,各实体的归属确定后,当然首先会观察分类的结果是否与预期相符,是否需要改变分类的类数,或重新指定初始类中心的位置后,再重新分类。此外还可以计算最终类中心之间的距离,以便分析哪些类之间关系接近,哪些类之间关系疏远。也可以对变量作一元方差分析,观察各变量在当前分类中的作用。有时会发现某些变量在当前的分类中不起多大作用,排除掉这类变量,也许能使分类结果与预期结果相符更理想。

14.7.2 K 均值分类方法应用实例

本章前面 14.5 节曾用等级聚类对殷墟颅骨进行了分类研究,其基础数据是表 14-7

所列22组颅骨的21项颅骨测量数据。图14-6显示了均值聚类的结果。22组颅骨分为3组。第1组包含爱斯基摩，楚克奇，布列亚和现代蒙古人等全部北亚类型的人群，但专门挑选的其颅骨形态接近北亚人群的殷墟祭祀坑I和中小墓II也进入此组。第2组由除祭祀坑I和中小墓II外的其他殷墟颅骨和其他东亚类型颅骨组成。第3组只包括高加索人一个组。现用SPSS软件的K均值分类程序对22组颅骨进行分类，也要求把它们分成3组或3类。数据按标准差标准化，初始类中心由计算机指定，迭代过程中采用running means方法。K均值分类程序的实际执行在经过18次迭代后停止。其分类结果由表14-9所示，可见与上述均值聚类的结果完全一致。9组北亚类型的颅骨加上殷墟祭祀坑I和中小墓II为第1类，高加索类型颅骨为第2类，全部东亚类型颅骨包括除祭祀坑I和中小墓II外的其他殷墟颅骨为第3类。说明分类结果是比较稳定的，聚类分析、主成分分析和K均值分类给出相同的分类结果。

SPSS的K均值分类程序的输出除给出最终的分类结果外，还给出每个实体到最终类中心的距离，初始和最终类中心的坐标值，18次迭代过程中每次类中心位置的变化量，每个变量的一元方差分析表，最终类中心之间的距离等信息。鉴于篇幅，这里我们仅列出实体分类结果和实体离最终类中心的距离表（表14-9）和最终类中心之间的距离表（表14-10）。

表14-9 K均值法对22组颅骨的分类结果和实体离分类中心的距离

颅骨编号	颅骨名称	分类	距离
1	殷祭祀坑Ⅰ	1	3.473
2	殷祭祀坑Ⅱ	3	3.197
3	殷祭祀坑Ⅲ	3	4.939
4	殷祭祀坑Ⅳ	3	1.770
5	殷祭祀坑Ⅴ	3	2.668
6	殷中小墓Ⅰ	3	1.949
7	殷中小墓Ⅱ	1	4.518
8	史前华北	3	3.457
9	柳　湾	3	4.577
10	仰韶（合并）	3	4.504
11	现代华北	3	4.197
12	现代华南	3	3.622
13	现代蒙古	1	2.982
14	爱斯基摩东南Ⅰ	1	2.887
15	爱斯基摩那俄康Ⅱ	1	3.877
16	爱斯基摩近代Ⅲ	1	2.777
17	楚克奇滨河Ⅰ	1	2.593
18	楚克奇驯鹿Ⅱ	1	2.889
19	布利亚西Ⅰ	1	1.986
20	布利亚东Ⅱ	1	3.019
21	布利亚外贝加尔Ⅲ	1	4.644
22	高加索	2	0.000

从上表看到，高加索组颅骨单独分为一类，实体到类中心的距离当然是 0，殷祭祀坑Ⅰ和殷中小墓Ⅱ虽分在北亚类，但相对而言它们离北亚组类中心的距离较远。

表 14-10 最终类中心间的距离

分类	北亚组	高加索组	东亚组 3
北亚组		6.564	5.420
高加索组	6.564		6.272
东亚组	5.420	6.272	

表 14-10 显示，(1)三个类中心相互间的距离大于各组颅骨到各自类中心的距离和(2)相对而言，北亚和东亚类中心间的距离小于它们各自到高加索组类中心的距离。

变量的一元方差分析表明在 21 项颅骨测量指数中，鼻根点角，面角，颅宽指数，鼻宽，眶宽，鼻指数对分类的影响相对较小。这里一元方差分析中的组间均方差和组内均方差之比，仅给出相应变量在分类中作用的大小，而没有统计学中用于显著性检验的意义。

最后需要说明，上面的分类过程选择了 running means 方法。如果在实体的迭代归类过程中不改变类中心的位置，那么在上面的例子中，仅经过 3 次迭代后数据就收敛了。得到的分类结果也不完全相同，与 4.5 节等级聚类的结果有差别。

14.8 模糊聚类简单介绍*

在根据古瓷的化学组成对其产地的溯源研究中，我国有的研究者使用模糊聚类方法(见苗建民等[1993]和李国霞等[2002])。而模糊聚类与本章 14.4 介绍的等级聚类或系统聚类在聚类思路方面有较大的差异。因此这里对模糊聚类作简要地介绍。

模糊聚类属于模糊数学内容。自 1965 年美国加州大学的 L.A.Zadeh 首先提出模糊集合的概念以来，模糊数学已发展为数学的一个重要分支。模糊数学是处理自然界和人类社会中大量难以用精确的数值变量描述的现象。例如一个篮球运动员出手投篮，我们不用也不可能测量球的速度和角度，目测就能以很高的概率正确预测球能否进入篮框。这类现象称为模糊现象。

模糊聚类也是对样本中的实体进行分类的过程，但它是通过实体间的模糊关系来进行聚类的。下面通过一个简单的实例来说明模糊聚类的过程，该例子引用自楼世博等编著的《模糊数学》(1983)。

假设有 5 个实体 $(x_1, x_2, x_3, x_4, x_5,)$，并且已经建立了它们之间相似系数矩阵(14-21)。

$$\boldsymbol{R}_1 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{vmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ 1 & 0.8 & 0 & 0.1 & 0.2 \\ 0.8 & 1 & 0.4 & 0 & 0.9 \\ 0 & 0.4 & 1 & 0 & 0 \\ 0.1 & 0 & 0 & 1 & 0.3 \\ 0.2 & 0.9 & 0 & 0.3 & 1 \end{vmatrix} \tag{14-21}$$

R_1 中的相似系数可以是根据描述实体的数值变量计算而得的,例如使用两个实体向量的夹角的余弦,也可以是根据模糊认识估计的。但是规定相似系数在0与1变动。在模糊数学中把相似系数看作是两个实体 x_i 和 x_j 间的模糊关系 $\mu_{R_1}(x_i,x_j)$,如果模糊关系十分密切,则 $\mu_{R_1}(x_i,x_j)$ 接近 1,反之,则接近 0。相似系数矩阵 R_1 在模糊数学中相应称为模糊关系矩阵。

对于矩阵的元素,

$$\text{当 } i = j \text{ 时}, \mu_{R_1}(x_i,x_j) = 1 \tag{14-22}$$

即主对角线上的元素等于 1。说明每个实体自己对自己的关系最接近,这称为模糊关系的自返性。此外对于 R_1 的元素显然还存在对称性关系,即有

$$\mu_{R_1}(x_i,x_j) = \mu_{R_1}(x_j,x_i) \tag{14-23}$$

因为模糊关系矩阵 R_1 具有自返性和对称性,R_1 又称为模糊相容矩阵。

直接利用模糊相容矩阵是不能进行实体聚类的,因为这个矩阵的模糊关系间不存在传递性。模糊关系的传递性要求,如果甲与乙聚一类和乙与丙聚一类,那么甲与丙也应该是一类的。对于公式(14-21)所示的模糊相容矩阵这种关系并不成立。例如规定 R_1 中模糊关系大于等于 0.8 的实体聚一类,那么 (x_1,x_2) 和 (x_2,x_5) 聚为一类,因为 $\mu_{R_1}(x_1,x_2) = 0.8 \geqslant 0.8$ 和 $\mu_{R_1}(x_2,x_5) = 0.9 > 0.8$。考虑到关系的传递性,$x_1$ 和 x_5 也应该属于同一类的。但是实际上 $\mu_{R_1}(x_1,x_5) = 0.4 < 0.8$,$x_1$ 和 x_5 并不属于同一类。这样就产生了矛盾。其实在 14.4 节一般的聚类过程中,我们也不是直接由相关系数矩阵完成聚类的,而是通过简单连接聚类,或均值聚类等方法来完成聚类过程的。

在模糊数学中是通过将 R_1 反映的模糊相容关系,转化为模糊等价关系来完成聚类过程的。转化的关键在于,在计算 x_i 和 x_j 的关系时,不仅考虑它们之间的直接关系,也要同时考虑 x_i 和 x_j 与其他实体的间接关系。这个转化是通过模糊关系矩阵的相乘 $R_2 = R_1 \circ R_1$ 来实现的。"∘"表示两个模糊矩阵相乘,这里我们不拟写出模糊矩阵相乘的一般公式,而直接写出二级模糊矩阵 R_2 的结果,并略作说明。

$$R_2 = \begin{vmatrix} 1 & 0.8 & 0.4 & 0.2 & 0.8 \\ 0.8 & 1 & 0.4 & 0.5 & 0.9 \\ 0.4 & 0.4 & 1 & 0 & 0.4 \\ 0.2 & 0.5 & 0 & 1 & 0.5 \\ 0.8 & 0.8 & 0.4 & 0.5 & 1 \end{vmatrix}$$

下面给出计算 R_2 的第一行第三列元素 $\mu_{R_2}(x_1,x_3)$ 的过程,说明为什么它等于 0.4。在 R_1 中 x_1 和 x_3 之间存在 5 对模糊关系系数,其中有 x_1 和 x_3 之间的直接关系,也有通过其他实体的间接关系。从每对关系的 2 个关系系数中选择数值小的系数。它们分别是

$$i = 1 \rightarrow \min[\mu_{R_1}(x_1,x_1),\mu_{R_1}(x_1,x_3)] = \min(1,0) = 0$$

$$i = 2 \rightarrow \min[\mu_{R_1}(x_1,x_2),\mu_{R_1}(x_2,x_3)] = \min(0.8,0.4) = 0.4$$

$$i = 3 \rightarrow \min[\mu_{R_1}(x_1,x_3),\mu_{R_1}(x_3,x_3)] = \min(0,1) = 0$$

$$i = 4 \rightarrow \min[\mu_{R_1}(x_1,x_4),\mu_{R_1}(x_4,x_3)] = \min(0.1,0) = 0$$

$$i = 5 \rightarrow \min[\mu_{R_1}(x_1,x_5),\mu_{R_1}(x_5,x_3)] = \min(0.2,0) = 0$$

然后从这 5 个数中选择最大的，得到 max (0, 0.4, 0, 0, 0) = 0.4。用同样的方法计算 $\boldsymbol{R}_2$ 的所有元素。不难证明 $\boldsymbol{R}_2$ 的每个元素都大于或等于 $\boldsymbol{R}_1$ 的相应的元素，即有

$$\mu_{R_2}(x_i, x_j) \geqslant \mu_{R_1}(x_i, x_j) \tag{14-24a}$$

接着还可以继续计算 $\boldsymbol{R}_3 = \boldsymbol{R}_2 \circ \boldsymbol{R}_2$（$\boldsymbol{R}_3$ 诸元素的数值不列出）和计算 $\boldsymbol{R}_4$

$$\boldsymbol{R}_4 = \boldsymbol{R}_3 \circ \boldsymbol{R}_3 = \begin{vmatrix} 1 & 0.8 & 0.4 & 0.5 & 0.8 \\ 0.8 & 1 & 0.4 & 0.5 & 0.9 \\ 0.4 & 0.4 & 1 & 0.4 & 0.4 \\ 0.5 & 0.5 & 0.4 & 1 & 0.5 \\ 0.8 & 0.9 & 0.4 & 0.5 & 1 \end{vmatrix}$$

公式(14-24a)可以推广到一般情况，即有对于任何的 $m < n - 1$，有

$$\mu_{R(m+1)}(x_i, x_j) \geqslant \mu_{Rm}(x_i, x_j) \tag{14-24b}$$

另外，还可以证明：(1)如果 $\boldsymbol{R}_1$ 是 $(n \times n)$ 的矩阵，n 为样本中实体的数目，那么当 $m = (n - 1)$ 时，

$$\mu_{R(n-1)}(x_i, x_j) = \mu_{Rn}(x_i, x_j) = \mu_{R(n+k)}(x_i, x_j) \tag{14-25}$$

式中 k 为任意的正整数。即 $(n - 1)$ 级模糊相容矩阵 $\boldsymbol{R}_{n-1}$ 的平方就等于 $\boldsymbol{R}_{n-1}$ 自己。(2) $\boldsymbol{R}_{n-1}$ 不仅保留自返性和对称性，而且必然具有传递性。因此 $\boldsymbol{R}_{n-1}$ 反映模糊等价关系，是模糊等价矩阵。这样可以直接使用 $\boldsymbol{R}_{n-1}$ 对实体聚类。

在上面的例子中，共有 5 个实体，$n = 5$，因此 $\boldsymbol{R}_4$ 已是模糊等价矩阵，可直接用于聚类。定义 $0 \leqslant \lambda \leqslant 1$，对于样本中的任意两个实体 x_i 和 x_j，只要 $\mu_{R4}(x_i, x_j) \geqslant \lambda$，$x_i$ 和 x_j 就聚为一类。

如果选 $\lambda = 0.9$，那么 (x_2, x_5) 聚类，其他实体各自为类，实体分成 4 类。

如果选 $\lambda = 0.8$，那么 (x_1, x_2, x_5) 聚类，其他实体各自为类，实体分成 3 类。

如果选 $\lambda = 0.5$，那么 (x_1, x_2, x_4, x_5) 聚类，实体 x_3 自成为一类，实体分成 2 类。

如果选 $\lambda = 0.4$，那么所有的实体均聚合为一类。

可见聚类的组数取决于 λ 值的选择。模糊聚类的结果也可以用树枝状聚类图来表示，如图 14-11 所示。

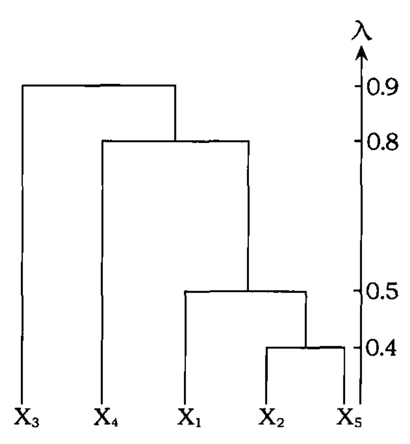


图 14-11 模糊聚类结果的树枝状图

从上面的讨论可见,对比模糊聚类和14.4介绍的等级聚类,两者的聚类策略是不同的,后者是一个逐步聚类的过程。在14.4.3节中,我们曾指出等级聚类的一个缺点,即后面的聚类过程会受到前面聚类实体的影响,如果某个实体因为某种偶然因素被错误地分到某一组,那么这个错误会自始至终地影响后面的聚类过程,有时会导致严重的后果。模糊聚类是可以避免这类问题的,因为它从开始就考虑了其他实体对每一对实体间关系的间接影响。等级聚类的另一重要缺点是不能显示原始变量对分类结果的作用,模糊聚类并不能免除这个缺点。

目前我国学者应用模糊聚类于瓷器的产地溯源研究,但是瓷器的化学组成是一些直接测量、相对精确的数值变量。考古器物的形态描述包含一系列模糊变量,应用模糊聚类于器物的分型定式也许是值得尝试的。可惜文献中未见到这方面的研究,可能是因为模糊分析的软件还没有像多元统计分析的软件那样普及。

模糊数学中还发展了所谓"软划分"和"硬划分"的分类方法,相应与14.6节的单元等级分划和14.7节的非等级的K均值分类方法有相似的分类思路。此外模糊模式识别对实体的归类,其所处理的问题与第十五章的多元判别分析是相同的。鉴于考古学研究对象的特征具有明显的模糊性,模糊数学在考古资料的定量研究中应该是很有前景的。关键之一也许是有关的模糊数学应用软件的开发和普及。

第十五章　判别分析与实体的归类

判别分析是一种对实体进行归类的多元统计方法。例如一个病人胸部透视照片上发现了阴影,要判断他患有结核或肿瘤哪种可能性大。为此要根据相当数量结核和肿瘤病人的资料,包括他们的胸片上阴影的位置、大小、形状、边缘的光滑度、病人的年龄、是否有低烧等多种指标来判断该病人患结核或肿瘤哪种可能性大,应该归入哪一个总体。在古瓷的鉴定方面,如果已知明代,清代景德镇官窑和现代青花瓷器的元素组成,对于一个未知来源的瓷器,我们希望根据测量其元素组成的来判断它是明、清的古瓷,还是现代的瓷器。判别分析就是根据实体的特征指标判断个体归属于哪种已知类型的一种方法。这里分别是在结核和肿瘤两类中选择,或者判断未知的青花瓷应属清瓷,明瓷和现代瓷等三种类型中的哪种,分别称为两总体和多总体的判别分析。在我国最早应用判别分析于考古研究的是王令红(1987),他将我国华南人、华北人、北亚蒙古人和波利尼西亚人的颅骨作为已知的类型,通过一系列的颅骨测量性状,判别在上述4类人群中,日本人与哪一人群最接近。判别分析的结论是:日本从最早期的港川人、绳文时代人到现代日本人都与同时期的华南人有最接近的亲缘关系。作为多元统计分析的判别分析的计算工作量很大,都是使用统计软件来完成的,因此后面的讨论,特别是实例应用部分,将结合SPSS软件的使用来进行。本章首先介绍判别分析的基本原理,然后15.2—15.4节讨论判别分析的3种方法,这3节的内容涉及矩阵运算等数学方法,对这方面内容不十分熟悉的读者可以不阅读这3节,而在了解了基本原理后直接阅读15.5节及后面的应用实例。在15.5.3小节中讨论了判别分析应用中的几个具体问题,希望能引起读者的注意。本章最后将简单介绍人工神经网络方法于实体的归类。

15.1　判别分析的基本原理

判别分析的基本思想可以通过图15-1来表示。

图15-1的例子代表一个种最简单的情况。实体的先验分类仅为A和B两类,而每个实体只需 x_1 和 x_2 两个变量来描述。图上分别用“×”和“∘”表示样本A和B中实体,显示了两个样本中实体的分布范围。理想情况下希望两个样本的实体点的分布接近正态分布,而且分布有相近的离散程度。现在执行一个线性的变换 $z = a_1x_1 + a_1x_2$,对每个实体 i 根据其原始坐标值(x_{i1}, x_{i2}),都可以计算一个 z_i 值,我们希望根据 z_i 值单个变量的大小来判断该实体应归属于两组中的哪一组。选择不同的 a_1 和 a_2 会得到不同 z_i 值。从图上可见,如果这样选择 a_1 和 a_2,使得计算所得的 z 值正好和图上的 z 轴相符,显然这时 z_i 值能够最佳地判别某个实体应该归属于 A 或者 B。选择别的 a_1 和 a_2 就不可能进行这样有效的判别。判别分析就是要寻找这样一个能对实体进行最佳归类的函数 z。z 称为判别函数,a_1 和 a_2 是判别函数的系数,z_i 称为实体 i 的判别得分。但是在图上有一块两个样本相互交叉重

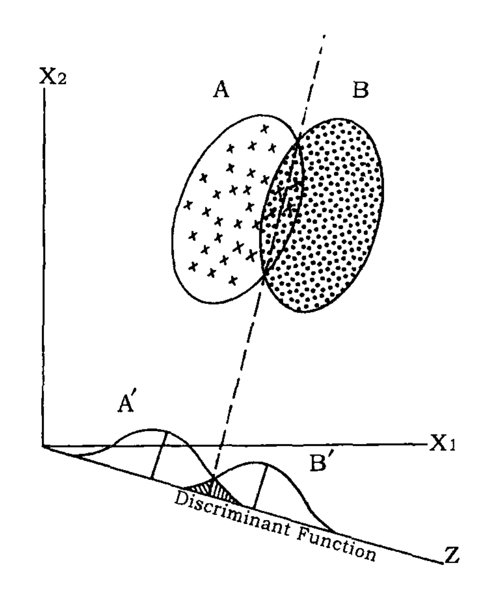


图 15-1 两个先验组实体判别分析原理的图示(引自 Joseph,1979)

叠的区域,如果实体及其判别得分 z_i 值正好处在这个区域,就有可能发生误判,误判率的大小是评价判别函数有效性的标准之一。此外,从图上还可以看到 x_1 对于判别函数 z 的贡献要比 x_2 大些。从上述关于判别分析的基本原理可知判别分析有三方面的任务和内容。

(一) 建立判别函数和实体的归类。

根据类属已知实体的坐标 x_{ij} 来计算推导判别函数,并根据判别函数对每个实体归类,包括对已知类属和未知类属的全部实体进行归类。这是三方面任务中最主要的。

1. 建立判别函数和判别实体归类的方法有多种,本章的 15-2、15-3 和 15-4 三节将分别介绍费舍判别方法、距离判别方法和贝叶斯概率判别方法等三种最常用的方法,也是 SPSS 软件的默认情况。它们都是原始分析变量 x_j 的线性函数。费舍判别是基于方差分析的原理,它对已知类属的实体分组,使得组内的方差尽量小而组间的方差尽量大。距离判别是计算实体与各先验组的中心之间的马氏距离,然后将每个实体分到离其最近的先验组中。贝叶斯判别则利用贝叶斯概率公式,在要求每个实体归属某类的概率最大和错判损失最小的条件下进行实体判别归类。因此贝叶斯方法考虑了各总体出现的概率 $P(G_i)$ 可能不相等,此外它还可以考虑错判造成的损失程度。但三种判别方法间不是完全孤立的,在一定条件下它们间是可以互相转换的。例如在先验组的数目是两个时,马氏距离判别函数等于第一和第二费舍判别函数的差值。如果先验组的数目是两个,而且两个总体均服从正态分布且出现的先验概率相等和两种错判所造成的损失也相等时,贝叶斯判别与距离判别等价,可以说距离判别是贝叶斯判别的特殊情况。

2. 需要说明,判别分析中判别函数的数目总是比分类数少一。即分 2 组时仅需建立 1 个判别函数,而分 n 组时需建立$(n-1)$个判别函数。但是各判别函数对判别归类的贡献是不相等的,往往只需考虑特征值最大的一、二个判别函数就能进行有效的判别归类。这点我们将在实例应用中看到。

3. 实际建立判别函数又可以有 2 种过程或 2 种方法,即变量的全选方法和逐步筛选法。全选方法又称全模型法,它是把全部变量($x_1, x_2, \cdots x_m$)一起引入判别函数。但是在实际的应用中并不是每个变量对于判别过程都起作用,某些变量不仅不能提高判别效果,反而会抑制其他变量的作用。另外变量之间的相关性也可能导致判别函数的不稳定和判别效果的降低。因此发展了逐步筛选法,逐步筛选法是根据一定的判据依次将对判别模型贡献最大的变量引入判别函数,同时剔除对判别模型影响不大的变量。最终仅部分原始分析变量进入判别函数,但其判别效果往往更好。

(二) 检验判别函数的有效性。

图 15-1 的例子中已经显示,判别分析中可能出现错判,即把原本属于 A 组的实体归类到 B 组,或者把原本属于 B 组的实体归类到 A 组。对于错判程度,可以用"判对率"来定量衡量。如果已知类属的实体数为 n,其中有 k 个实体归类判别正确,那么判对率为 $\frac{k}{n}$。但是这样计算的判对率对判别函数有效性的估计是偏高的,因为判别函数本身是根据先验分组的数据建立的。由此发展了一种称为"leave one out"的方法,它每次将一个实体排除在外计算判别函数,用这个判别函数计算被排除实体的判别得分并进行归类。再计算"leave one out"条件下的判对率$\frac{k}{n}$。一般后者要比前者为小,但更实际地估计判别函数的有效性。判别函数的有效性也可通过 Wilk's λ 值来表示,Wilk's λ 值的定义在 15.5 节中讨论。

(三) 分析原始分析变量对判别函数的贡献。

在判别分析的实际应用中,最终希望能解释为什么某个实体被判归属于某类,这就需要了解原始分析变量对判别函数的贡献。每个分析变量对于判别函数的贡献是不同的,对于图 15-1 的例子,直观上变量 x_1 对于判别函数贡献较 x_2 为大。对于多变量的情况,判别函数 F 的形式如后面的式 15-2 所示。如果原始数据 x_{ij} 已作了标准化转换,那么所得判别函数系数 a_i 的数值正比于变量 x_i 对于判别函数的贡献。此外变量 x_i 与判别得分间的简单相关系数也反映了变量 x_i 对于判 别函数的贡献。SPSS 软件在执行判别分析程序时,将输出一个 ($n \times 1$) 的结构矩阵,该矩阵的元素是判别得分与各变量间的简单相关系数,并按数值的大小排列。因此结构矩阵中的元素的排列次序反映了变量对判别函数贡献大小的次序。我们将在实例应用节再回到这个问题。

后面的 3 节将介绍费舍判别等三种方法。需要说明这里我们均限于讨论两总体的情况,即已知先验组的数目是 2 组。关于多总体的情况,仅将在 15.7 和 15.8 节结合实例应用予以讨论。对阅读这 3 节有困难的读者,可以跳过这些内容,直接阅读 15.5 节的实例应用。

15.2 费舍判别方法*

假设有 n 个实体分别属于 A, B 两个样本,它们各有 n_1 和 n_2 个实体,每个实体均被 m 个变量描述,x_{ij} 表示第 i 个实体的第 j 个变量的取值并且是已知的。进行判别分析有两个假设前提。(1)这两个样本均来自正态分布总体和(2)两总体的协方差矩阵相等,即有

$S_1 = S_2$。样本的协方差矩阵 S 是 $m \times m$ 的矩阵,其第 i 行第 j 列元素 s_{ij} 的定义是

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \tag{15-1}$$

$\bar{x}_i$ 和 $\bar{x}_j$ 是变量 x_i 和 x_j 的组平均值,协方差矩阵对角线上的元素就是方差。希望建立的判别函数的形式为

$$F = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m \tag{15-2}$$

式中的 诸 a_i 值称为判别函数的系数。对于每个实体都可以计算一个判别得分值 f_i,而对每个样本可以计算其判别得分的平均值,即 $\bar{f}_1$ 和 $\bar{f}_2$,和判别得分的方差 s_1^2 和 s_2^2。费舍判别认为最佳的判别是希望 $\bar{f}_1$ 和 $\bar{f}_2$ 相差尽量大,而组内的方差 s_1^2 和 s_2^2 尽量小。即组内的各实体聚集得尽可能密集,而两个组中心间的距离尽可能远。应该根据上述的要求来确定判别函数的诸系数。

因为已经假设两个总体的协方差矩阵相等,可以计算 s_1^2 和 s_2^2 间的平均值 s_w^2,称为计权平均组内方差。

$$s_w^2 = \frac{[(n_1-1)s_1^2 + (n_2-1)s_2^2]}{(n_1+n_2-2)} \tag{15-3}$$

组内方差 s_w^2 反映组内实体判别得分的离散程度。同样可计算组间的方差 s_b^2:

$$s_b^2 = n_1(\bar{f}_1 - \bar{f})^2 + n_2(\bar{f}_2 - \bar{f})^2 \tag{15-4}$$

s_b^2 反映两个样本各自平均判别得分的离散程度。应该这样来确定公式(15-2) 中的诸 a_i 值,使得组间方差相对于组内方差的比值 $\frac{s_b^2}{s_w^2}$ 尽量大。上述判别分析的基本思想和计算判别函数的系数的方法是费舍首先提出来的,因此称为费舍判别方法,他借用了一元方差分析的思想(见第七章公式 7-16)。至于如何根据使 $\frac{s_b^2}{s_w^2}$ 尽量大的原则来具体确定诸 a_i 值,在数学计算上是较复杂的,这里不可能作详细的讨论,而是直接给出结论。可以证明,满足上述要求的判别函数的系数矢量 $\boldsymbol{a}$,($\boldsymbol{a}$ 的转置矢量是 $\boldsymbol{a}' = (a_1, a_1, \cdots a_m)$) 正比于

$$\boldsymbol{S}_w^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \tag{15-5}$$

由此可以计算式(15-2) 的诸 a_i 值。式中的 $\boldsymbol{S}_w$ 是组内计权平均协方差矩阵,$\bar{\boldsymbol{x}}_1$ 和 $\bar{\boldsymbol{x}}_2$ 分别是表示两组中心坐标的矢量。由此可以看出,两组平均判别得分的差值($\bar{f}_1 - \bar{f}_2$) 就是在原来的变量坐标空间中两组中心间马氏距离的平方,即

$$(\bar{f}_1 - \bar{f}_2) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_w^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_2) = D^2 \tag{15-6}$$

而实体 i 到第一组中心的马氏距离平方 D_{i1}^2 为

$$D_{i1}^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_1)'\boldsymbol{S}_w^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_1) \tag{15-7}$$

$\boldsymbol{x}_i$ 是代表第 i 个实体空间位置的矢量。式(15-7)也可改写成

$$D_{i1}^2 = -2(\bar{\boldsymbol{x}}_1'S_w^{-1}\boldsymbol{x}_i - c_1) + c \tag{15-8}$$

式中 c_1 对组 1 是一个常数,c 是与组别无关的常数。括弧内的公式是 x_i 的线性函数,称为第一组费舍线性判别函数。同样方法计算实体 i 到第二组中心的马氏距离平方 D_{i2}^2 为

$$D_{i2}^2 = -2(\bar{\boldsymbol{x}}_2'S_w^{-1}\boldsymbol{x}_i - c_2) + c \tag{15-9}$$

式(15-9)括弧内的内容是第二组的费舍线性判别函数。对于任何一个实体既可以根据费舍线性判别函数的大或小,也可以根据马氏距离的近或远,归类到两组中的哪一组。我们将在 15.3 详细讨论怎样利用马氏距离进行判别。

15.3 距离判别方法*

上一节讨论了费舍判别方法,并在讨论中进一步阐述了判别分析的基本思想。本节将介绍距离判别方法。距离判别的思想也是简单明了的。在图 15-1 上,A,B 两组实体都有一个中心,考察任何一个代表实体的点 (x_{i1}, x_{i2}),它离哪个中心近,距离短,这个实体就应该归属哪一组。但是使用我们通常熟悉的欧氏距离作为距离的度量不合适,因为当测量变量的单位变化时,距离的数值相应发生变化,而且当几个变量间存在较强的相关关系时,也会影响距离的度量。所以需要使用马氏距离作为实体间和实体和组中心间距离的度量。在 14.3.1 小节中已简单介绍了马氏距离的概念(见公式(14-10))。空间任何两点 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 的马氏距离平方为

$$D^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)'\boldsymbol{S}^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2) \tag{15-10}$$

$\boldsymbol{S}$ 为协方差矩阵,要求 $\boldsymbol{S} > 0$。为了书写的方便,组中心的坐标 $\bar{\boldsymbol{x}}_i$ 写为 $\boldsymbol{\mu}_i$。假设对于两个样本有 $\boldsymbol{S}_1 = \boldsymbol{S}_2 = \boldsymbol{S}$,即它们的协方差矩阵相等,那么任何样品点 $\boldsymbol{x}$ 到两个组中心的马氏距离平方分别为

$$D^2(\boldsymbol{x}, \boldsymbol{\mu}_1) = (\boldsymbol{x} - \boldsymbol{\mu}_1)'\boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) \tag{15-11a}$$

$$D^2(\boldsymbol{x}, \boldsymbol{\mu}_2) = (\boldsymbol{x} - \boldsymbol{\mu}_2)'\boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \tag{15-11b}$$

计算两个马氏距离平方的差

$$\triangle D^2 = D^2(\boldsymbol{x}, \boldsymbol{\mu}_2) - D^2(\boldsymbol{x}, \boldsymbol{\mu}_1) \tag{15-12}$$

如果 $\triangle D^2 > 0$,那么实体 x 归属第二组,反之如果 $\triangle D^2 < 0$,那么实体 x 归属第一组。在两个协方差矩阵相等的假设条件下,公式 15-12 经过一系列运算后可写成

$$\triangle D^2 = 2(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})'\boldsymbol{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 2(\boldsymbol{x} - \bar{\boldsymbol{\mu}})'\boldsymbol{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{15-13}$$

令 $\boldsymbol{A} = \boldsymbol{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$,和 $G(x) = \frac{\triangle D^2}{2}$,则式 15-13 可改写为

$$G(x) = (\boldsymbol{x} - \bar{\boldsymbol{\mu}})'\boldsymbol{A} = \boldsymbol{A}'(\boldsymbol{x} - \bar{\boldsymbol{\mu}}) \tag{15-14}$$

$G(x)$是 x 的线性判别函数,A 是判别函数中的诸 a_i 系数,即判别系数。对比式(15-8)和(15-9)可知,马氏距离判别函数等于第一和第二费舍判别函数的差。

15.4 贝叶斯概率判别方法*

前一节介绍的马氏距离判别,其思路明晰,涉及的计算过程相对简单而且结论明确。另外它实际上对总体的分布并没有什么前提要求,因此得到广泛的应用。但是它没有考虑各总体出现的概率 $P(G_i)$ 可能不相等,此外它默认错判造成的损失是常数,即认为 $C(i \mid j) = C(j \mid i)$。$C(i \mid j)$ 是本属于 G_j 组的实体误判归属 G_i 组所造成的损失。贝叶斯

判别方法正是为解决这些问题而提出的。下面作简要介绍。

设有 k 个总体 $G_1, G_2 \cdots\cdots G_k$，它们的分布密度函数为 $f_i(x)$，总体 G_i 出现的概率为 $P(G_i)$，且有 $\sum_i P(G_i) = 1$，$C(i \mid j)$ 是本属于 G_j 组的实体误判归属于 G_i 组所造成的损失。如果 $i \neq j$，则 $C(i \mid j) > 0$；如果 $i = j$，$C(i \mid j) = 0(i, j = 1, 2, \cdots\cdots k)$。贝叶斯判别的基本原则是要求每个实体归属某类的概率最大和错判损失最小。贝叶斯判别函数的建立需要关于 $P(G_i)$，$f_i(x)$ 和 $C(i \mid j)$ 的知识，而且计算十分复杂。对于较简单的情况 $k = 2$，即只有两个总体的情况，则有

$$\text{如果} \frac{f_1(x)}{f_2(x)} > \frac{C(1 \mid 2)P(G_2)}{C(2 \mid 1)P(G_1)}, \text{实体 } x \text{ 归属于第 1 组}, x \in G_1$$

$$\text{如果} \frac{f_1(x)}{f_2(x)} < \frac{C(1 \mid 2)P(G_2)}{C(2 \mid 1)P(G_1)}, \text{实体 } x \text{ 归属于第 2 组}, x \in G_2 \qquad (15\text{-}15)$$

如果 $P(G_1) = P(G_2)$，即如果两个总体出现的概率相等(称为先验概率相等)，而且 $C(1 \mid 2) = C(2 \mid 1)$，即两种误判所造成的损失也相等，则公式(15-15)的左边等于1，这样实体 x 归属于哪个总体，完全取决于分布密度函数 $f_1(x)$ 与 $f_2(x)$ 哪个大。最理想的情况是，除 $P(G_1) = P(G_2)$ 和 $C(1 \mid 2) = C(2 \mid 1)$ 外，而且 $f_1(x)$ 与 $f_2(x)$ 均服从正态分布，这时贝叶斯判别与距离判别完全等价，可以说距离判别是贝叶斯判别的特殊情况。

如果只是假设 $C(1 \mid 2) = C(2 \mid 1)$，那么实体 x 归属于 G_i 组的概率为

$$P(G_i \mid x) = \frac{P(x \mid G_i)P(G_i)}{\sum_i P(x \mid G_i)P(G_i)} \qquad (15\text{-}16)$$

这是我们熟悉的贝叶斯公式(见第四章公式(4-10))。式中的 $P(G_i)$ 是总体 G_i 出现的先验概率，也可以理解为实体属于 G_i 的先验概率。对于两总体的判别分析有两种选择来确定 $P(G_i)$，(1)认为每个总体出现的概率是相等的，即有 $P(G_1) = P(G_2) = 0.5$。(2)认为样本是总体的代表，以样本的相对容量作为相应总体 G_i 出现的先验概率，即有

$$P(G_1) = \frac{n_1}{n_1 + n_2} \text{ 和 } P(G_2) = \frac{n_2}{n_1 + n_2} \qquad (15\text{-}17)$$

公式中的 n_1 和 n_2 分别是样本 1 和样本 2 中的实体数目。

$P(x \mid G_i)$ 是实体 x 在总体 G_i 中出现的条件概率，其数值依赖于总体的密度分布函数 $f_i(x)$，如果已知 $f_i(x)$ 服从正态分布函数，而且其平均值和方差可以用样本的平均值和方差估计，那么 $P(x \mid G_i)$ 是可以计算得到的。从而利用公式(15-16)可以计算实体 x 分别归属于总体 G_1 和 G_2 的后验条件概率 $P(G_1 \mid x)$ 和 $P(G_2 \mid x)$，根据后验概率的大小确定实体 x 归属于两组中的哪一组。SPSS 软件中用户可以选择先验概率，条件概率是根据正态分布计算的，程序执行结果给出实体属于各总体的后验概率。

15.5 两总体全选模型判别分析的实例：殷墟颅骨的种系判别

本节和后面的 3 节将通过判别分析的应用实例，先后介绍两总体的全选模型和逐步筛选方法，以及多总体的全选模型和逐步筛选方法。

第十四章讨论等级聚类和非等级的 K－均值分划中都曾以北亚、东亚和殷墟的 22 组

颅骨的分类为例子。为了便于比较,本节也使用这 22 组颅骨的数据作判别分析。表 14-7 列出了这 22 组颅骨 21 项测量指标的平均值。第十四章进行的均值聚类,Ward's 方法聚类和 K-均值分划在将这批颅骨分为 2 类或 2 组时,得到的分类结果是相同的。第 1 组包括全部 9 组北亚类型的颅骨和专门挑选的形态接近北亚类型的殷墟祭祀坑 I 和殷墟中小墓 II 颅骨,共 11 组。第二组也是 11 组,由剔除了殷墟祭祀坑 I 和殷墟中小墓 II 的其他 10 组东亚类型颅骨加上高加索类型颅骨组成。如果要求进一步将 22 组颅骨分为 3 组时,均值聚类和 K-均值分划都将高加索类型颅骨从东亚组中分离而独立为一类,但 Ward's 方法聚类却是将北亚组继续分为典型北亚类型和极地北亚类型两个亚组,高加索类型颅骨仍保留在东亚组中(见图 14-6 和图 14-7)。

15.5.1　SPSS11.0 软件全选模型判别分析程序的对话框

下面使用 SPSS11.0 软件的全选法判别分析程序来检验上面的分类结果。在叙述判别分析的过程和结果时同时对 SPSS 中判别分析程序的选项和输出作说明。需要说明,前面曾多次提到判别分析的效率,但没有介绍怎样定量地估计判别效率,也没有讨论怎样检验判别的有效性,这方面的内容也将通过这个实例作讨论。

判别分析中暂时不考虑高加索颅骨,将北亚组 11 组颅骨(含殷墟祭祀坑 I 和殷墟中小墓 II)和东亚组 10 组颅骨作为两个已知组或先验组。为此在表 14-7 的 SPSS 的数据文件中要添加一个分类变量,对北亚组该变量取值为 1,对东亚组取值 2,对高加索颅骨取值 3。在判别分析程序的对话窗口中通过输入分类变量名和分类变量的取值,来选择进入分析阶段的实体;全部 21 项测量指标作为分析变量输入并选择全选法。打开"Classify"对话窗口:(1)选择两组的先验概率 $P(G_i)$ 相等。(2)要求计算和输出各实体的判别得分,它们与两组中心间的马氏距离,它们的归属组别和相应的概率。(3)选择使用组内协方差矩阵作分析。回到判别分析的主窗口,单击"OK",即可执行判别分析程序。

15.5.2　执行 SPSS11.0 软件全选模型判别分析程序的输出内容和解释

1. 程序首先列表汇总输出:输入的实体数目和进入分析阶段的实体数目(本例分别为 22 和 21),每个先验分组中的实体数目和全部分析变量的名称。

2. 程序显示因其容忍度太低而被程序自动排除在分析变量之外的变量名称。在本例中颅宽高指数、中上面角、鼻指数、眶指数和额宽指数等 5 个变量因其容忍度太低而被排除。变量的容忍度定义为 $1-R_i^2$,R_i 为变量 i 与其他所有变量的总线性相关系数。容忍度太低的变量对模型的贡献很小而且可能引起计算中的麻烦。

3. 程序给出 Wilk's λ 值并对判别函数作相应的显著性检验,如表 15-1 所示。Wilk's λ 值定义为实体判别得分的组内离差平方和与总平方和的比值,它是判别得分总变异中未能被组间差异所能解释部分的百分比。Wilk's λ 值总在 0—1 间变动,这个检验量越接近 0 表示未能被解释部分的比例越小,判别判别函数的有效性越高。在两个组判别得分的总体间不存在差异的零假设下,Wilk's λ 值可转化为一个近似服从 χ^2 分布的统计量,其自由度等于被保留的变量数。对于所分析的例子,$\lambda=0.013$,十分接近 0,χ^2 检验也拒绝两总体的均值间不存在差异的零假设。

表 15-1 Wilk's λ 值和相应的 χ^2 检验

Test of Function(s)	Wilk's Lambda	Chi-square	df	Sig.
1	0.013	47.541	16	0.000

4. 输出判别函数的系数。因为先验分组为两组，只生成一个判别函数。表 15-2a 和 15-2b 分别给出标准化和未标准化的判别函数系数 a_i，它们分别对应于标准化和未标准化的原始数据 x_{ij}，将它们代入公式(15-1) 就可得到判别函数。

表 15-2a 标准化判别函数系数

	Function
	1
颅 长	-0.358
颅 高	-2.071
颅 宽	-0.827
耳上颅高	2.366
最小额宽	-0.929
颧 宽	-0.791
上面高	1.051
鼻 高	-1.140
鼻 宽	0.203
眶 宽	0.760
眶 高	0.709
面 角	-0.356
齿槽点角	2.449
鼻根点角	-0.483
颅指数	1.270
颅长高指数	1.245

15-2b 非标准化判别函数系数

	Function
	1
颅 长	-0.189
颅 高	-0.558
颅 宽	-0.432
耳上颅高	1.561
最小额宽	-0.681
颧 宽	-0.356
上面高	0.564
鼻 高	-1.033
鼻 宽	0.140
眶 宽	0.596
眶 高	0.754

续表

	Function
	1
面 角	-0.180
齿槽点角	1.936
鼻根点角	-0.317
颅指数	0.527
颅长高指数	1.273
(Constant)	-173.525

非标准化的判别函数系数中有一个常数项。利用非标准化的判别函数可计算得到两组中心的判别得分,如表 15-3 所示:

表 15-3 各组中心的判别得分

	Function
Cluster Number of Case	1
1	-7.819
2	8.601

如果在程序执行前,在"Statistic"对话框中选择了"Fisher's"项,程序也可显示费舍函数相应的两列系数。

5. 程序输出的"Casewise Statistics"表显示了每一个实体的判别分析结果,包括每个实体的先验分组和判别分析归组,实体的判别得分,实体与两个组中心间的马氏距离,实体归属到两组中每一组的后验概率等。这是判别分析的重要结果,但鉴于该表格所占篇幅太大,这里不予列出。

程序也给出判别归类汇总表,它统计各实体的先验分组和判别归组是否符合,计算判别分析的判对率和误判率。在本节分析的实例中,全部 21 组颅骨的先验分组与判别归组均为一致,判对率达 100%。在前面判别分析的分析阶段时,高加索类型颅骨是被排除在外的,但当得到判别函数后,也可以计算高加索类型颅骨这类未进入分析阶段的实体的判别得分,并对它进行归组。高加索类型颅骨被归入第 2 组,即东亚组。与第十四章中两种聚类方法和 K-均值分划的分类结果是一致的。

实际上利用前面计算的判对率来估计判别函数的有效性,往往是估计过高,过分乐观的,因为判别函数是在考虑了先验分组的条件下计算得到的。为了更现实地计算判对率发展了一种称之为"Leave-one-out"的方法,它逐次将一个实体排除在外计算判别函数,用这样计算而得的判别函数计算被排除实体的判别得分并对它进行归类。对于 21 组颅骨,"Leave-one-out"判别分析的结果有 2 个实体被误判,分别是东亚的殷祭祀坑Ⅲ和现代华南颅骨,它们被归入北亚组,"Leave-one-out"判别分析的判对率为 90.5%。表 15-4 是归类结果汇总表,表的上半部分汇总 21 个进入分析阶段的实体的判别分析结果。该表的下半部分汇总"Cross-validated",即"leave one out"判别分析的结果,可见第 2 先验组中有 2 个实体误判,被归类进第 1 组。表中"ungrouped cases"指本例中的高加索类型颅骨。

表 15-4 全选判别方法归类结果汇总表

			Predicted Group 1	Predicted Group 2	Total
Original	Count	1	11	0	11
		2	0	10	10
		Ungrouped cases	0	1	1
	%	1	100.0	0.0	100.0
		2	0.0	100.0	100.0
		Ungrouped cases	0.0	100.0	100.0
Cross-validated	Count	1	11	0	11
		2	2	8	10
	%	1	100.0	0.0	100.0
		2	20.0	80.0	100.0

判别分析的结果也可以以图形的形式输出。图 15-2 是 22 组颅骨判别得分的直方图,可以看出北亚组颅骨分布在左边,而东亚组在右边,两者间的距离是拉开的,反映判别的有效性较高。前面提到判别分析程序对原始数据表中的每一个实体,包括其先验类属未知的实体都进行归类,将其归入某类组。未进入分析阶段的高加索组虽被归入了东亚组,但是高加索颅骨的判别得分为 2.89,与东亚组组中心的判别得分 8.601 相距颇远,这在图中也明显可见。因此判别分析对实体的归类并不是证明该实体一定属于所归属的类组,而仅表示在各先验组之间,该实体根据其属性的取值更接近于其所归属的类组。具体到高加索组颅骨的归类,应该理解为相对于北亚组,它更接近于东亚组,归入东亚组不是说明它一定属于东亚组。

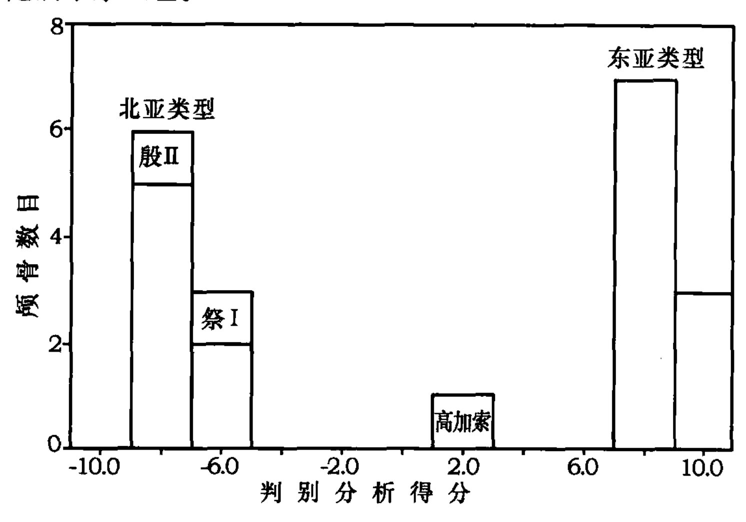


图 15-2 22 组颅骨判别得分的直方图

6. 最后,人们还希望了解是哪些原始变量决定了每个实体的归类,即希望了解各原始分析变量对判别函数贡献的大小。这可以根据标准化判别函数的系数来估计。表 15-2a 列出了标准化判别函数的各系数,从表中看到齿槽点角,耳上颅高和颅高等变量的判别系数

均大于 2,而鼻宽,面角等变量的判别系数很小,小于 0.3,因此前面 3 个变量对判别函数的贡献要显著大于鼻宽,面角等变量。此外 SPSS 判别分析程序的执行还输出一个结构矩阵,如表 15-5 所示。结构矩阵的元素是各原始分析变量与标准化判别函数值间的相关系数,并按数值的大小排列。相关系数反映了各变量对判别函数的贡献,因此表中排列靠前的变量对判别函数贡献较大。虽然矩阵排列靠前的变量和标准化判别函数中系数大的变量不全相同,但是它们都是反映区分蒙古大人种北亚和东亚类型颅骨的主要特征,如颧部面部的宽窄长短和颅部的相对高低。这些特征在决定判别函数中有很大的权重。

表 15-5　判别分析的结构矩阵(按数值大小排列)

	Function
	1
颧 宽	-0.218
颅长高指数	0.190
最小额宽	-0.176
眶指数*	-0.160
颅 高	0.145
眶 高	-0.140
颅宽高指数*	0.138
耳上颅高	0.125
上面高	-0.118
颅 宽	-0.117
鼻 高	-0.110
齿槽点角	0.088
颅指数	-0.082
中上面角*	-0.071
鼻指数*	-0.066
颅 长	-0.044
面 角	-0.033
眶 宽	-0.026
鼻根点角	0.025
鼻 宽	0.016
额宽指数*	0.015

“*”表示此变量未被应用于判别分析,即因容忍度低而被排除的变量。

7. SPSS 判别分析程序根据用户的选择,还可以输出其他的统计量,例如各个协方差矩阵,费舍判别函数的系数,对各个变量进行一元方差分析等。

15.5.3　判别分析中的几个问题

1. 关于正态分布与歧离实体。在本章的最初曾提到样本来自正态总体是判别分析的一个前提。但是在后面的讨论中可以看到这个前提要求并非是绝对严格的,虽然前提的成立使得贝叶斯判别与距离判别等价。判别分析实际使用的经验也表明,总体的实际分布略偏离正态不会严重影响判别分析的结果。但是偏离组中心很远的歧离实体(Outli-

er)的存在会影响判别函数的稳定性和有效性,而且这些歧离实体也破坏总体间协方差一致性的前提。因此在进行判别分析前,要对原始数据作前期观察,剔除掉严重偏离样本均值的歧离实体。

2. 总体间协方差的一致性问题。15.2—15.4 介绍的三种判别标准都要求总体间协方差的一致性。已发表的不少应用判别分析的研究论文中并没有严格地关心方差一致性前提,而直接注意判别效果的优劣。SPSS 的判别分析程序提供了检验方差一致性的 Box 方法。但 Box 检验对进入分析的实体的数目有一定要求,我们将在下一节逐步筛选判别分析方法中讨论。

3. 实体与变量的数目。用于建立判别函数的每个先验组中实体的数目不应太少,这不仅是 Box 检验的要求。如果实体数目太少,即样本容量太低会导致判别函数的不稳定性,建立判别函数的基础是已知先验分类的实体的数据。理想情况下要求 $n_i > 2m$, 即每个先验组实体的数目大于 2 倍的变量数目。当实际测量的实体的数量太少时,应考虑适当减少选择变量的数目。逐步选择判别比全选法的一个优点是它排除了一些对判别作用不大的变量。

4. 判别分析对每一个未知类属的实体归类,但这并不证明该实体就来自所归类属的总体,而仅仅表明在诸先验类属中,该实体的性状最接近于某个类属。正如本节讨论的实例中,判别分析将高加索类型颅骨判归蒙古人种东亚类型组,但是高加索类型颅骨并不属于蒙古人种东亚类型。判别分析的结果仅仅表明,相对于北亚类型,高加索类型颅骨更接近于东亚类型。这个结论不应往外延伸。近年我国有不少研究单位在建立各类古代名瓷的化学组成数据库,并试图根据这些数据库,使用判别分析的方法对市场上未知来源的瓷器进行鉴定,判别它们是否属于某类古代名瓷。鉴于判别分析上述的基本的特点,必须慎审对待这种瓷器鉴定方法,一般情况下辨伪较为容易,而希望确认某件瓷器确实属于某类古代名瓷应十分小心。

15.6 两总体逐步筛选模型判别分析的实例:殷墟颅骨种系的再判别

15.6.1 逐步筛选模型判别分析的思想和 SPSS 对话框

本节介绍逐步筛选判别分析方法。15.5 节讨论的全模型法把全部变量 ($x_1, x_2, \cdots\cdots x_m$) 一起引入判别函数。但是在实际的应用中并不是每个变量对于判别过程都起作用,某些变量不仅不能提高判别效果,反而会抑制其他变量的作用。另外变量之间的相关性也可能导致判别函数的不稳定和判别效果的降低,因此发展了逐步筛选法。逐步筛选法是根据一定的判据首先将一个对判别模型贡献最大的变量引入判别函数,第二步再将贡献次大的变量引入,这样一步步的筛选变量进入模型。同时新变量的进入可能会因为变量间的相关性而降低已选变量对判别模型的贡献,这样又要根据一定的判据检验是否需要将哪个已被选入的变量剔除,不断的筛选进入和不断的剔除,直到已选入的变量都符合判据的要求而模型外未选或被剔除的变量都不符合被选进入模型的判据时,逐步筛选的过程结束。逐步筛选法建立的判别函数仅包含部分变量,但是其判别效果往往更好。逐步筛选法的判据可以用贝叶斯判别函数,可以用马氏距离也可以用费舍判别标

准。关于判据的具体数值标准将在本节后面讨论。

讨论逐步筛选判别分析方法将依然使用 15.5 节中安阳殷墟 22 组颅骨的例子，分析中仍将北亚组 11 组颅骨（含殷墟祭祀坑 I 和殷墟中小墓 II）和东亚组 10 组颅骨作为两个已知组，暂时不考虑高加索颅骨。使用 SPSS 程序的分析过程是先打开 SPSS 判别分析的主对话窗口，输入分类变量和根据分类变量的取值选择进入分析阶段的实体（高加索颅骨不进入分析阶段），全部 21 项测量指标作为分析变量输入，这与执行全选方法的判别分析过程一致。在主对话框中选择"Use stepwise method"，这时对话框中"Method"键被激活。在"Method"对话框中有方法栏和临界值栏。方法栏内可选的方法有 Wilk's λ，未能被解释的方差，马氏距离等多种方法，SPSS 默认的是 Wilk's λ 方法。临界值栏内可选的有 F 值和 F 的概率值两项，默认的是 F 值，变量被选取和移出的 F 临界值分别规定为 3.84 和 2.71。建议在"Method"对话框中接受"Summary of steps"选项，以便观察每步的执行过程。由于逐步筛选方法最终选择的变量数少，有可能用 Box 方法检验总体间协方差的一致性，因此在"Statistic"对话框中可选择"Box's M"。"Classify"对话框中的各选项与全选模型是相同的，也选择两组的先验概率 $P(G_i)$ 相等。

15.6.2　SPSS 程序执行两总体逐步筛选模型判别分析的输出

1. 逐步筛选判别分析程序的输出与全选程序相似，首先也是汇总并列表显示输入的实体数目、用于判别分析的实体数目、每个先验分组中的实体数目和全部分析变量的名称等。

2. 因为要求作 Box 检验，程序的执行将输出对总体间协方差一致性的 Box 检验结果（表 15-6）。对于 22 组颅骨的实例，统计量 Box's M = 18.335，相应的 F 检验的显著性水平是 0.171。由此在显著性水平为 0.171 的条件下，Box 检验接受关于北亚和东亚两组颅骨数据，其总体的协方差无显著差异的假设。

表 15-6　总体间协方差一致性的 Box's M 检验

Box's M		18.335
F	Approx.	1.406
	df1	10
	df2	1683.537
	Sig.	0.171

3. 逐步判别过程第一步计算所有变量的 Wilk's λ 值，选择其 Wilk's λ 值最小的变量进入模型，因为 Wilk's λ 值是组内平方和与总平方和的比值。殷墟颅骨的实例中颧宽的 Wilk's λ 值最小，为 0.221，而且根据其 Wilk's λ 值和自由度，计算所得的 F 值等于 66.87，大于 3.84（见表 15-7 和表 15-9），因此变量颧宽首先被选进入模型。第二步是在颧宽已选的条件下，再计算其他所有变量的 Wilk's λ 值，变量耳上颅高被选进入模型，因为其 Wilk's λ 值最小为 0.84，$F > 3.84$。这样一步步选择变量，相继最小额宽和齿槽点角被选。当第四步齿槽点角被选后，其他变量根据 Wilk's λ 值计算的 F 值均小于 3.84，因此程序执行终止。此外在程序执行中每个新变量的进入，会改变先进入模型的变量的 Wilk's λ 值和 F 值，如果某个变量的 F 值小于 2.71，则该变量将被从模型中剔除。在本例中没有发生已选变量被剔除的情

况。SPSS 程序分别列表(表 15-7,表 15-8)给出已选变量和未选变量的容忍度,Wilk's λ 值和 F 值,清楚地显示每一步当一个新变量被选后,哪个变量应被继续选入和哪个已选变量应被剔除。鉴于未选变量表的篇幅较大,这里仅列出变量选择的第三步,即当颧宽,耳上颅高和最小颅宽 3 变量已被选后,其他 18 个未被选变量的情况(表 15-8)。

表 15-7 逐步筛选判别分析程序执行过程中各步被选的分析变量表

Step		Tolerance	F to Remove	Wilk's Lambda
1	颧宽	1.000	66.874	
2	颧宽	0.645	82.060	0.464
	耳上颅高	0.645	29.691	0.221
3	颧宽	0.574	37.530	0.128
	耳上颅高	0.432	48.996	0.155
	最小颅宽	0.666	18.583	0.084
4	颧宽	0.429	49.337	0.127
	耳上颅高	0.395	47.652	0.124
	最小颅宽	0.658	8.238	0.047
	齿槽点角	0.690	4.467	0.040

表 15-8 逐步筛选判别分析程序执行过程第 3 步后未被选的分析变量表

Step		Tolerance	Min. Tolerance	F to Enter	Wilk's Lambda
3	颅 长	0.727	0.431	0.437	0.039
	颅 宽	0.760	0.425	0.222	0.039
	颅 高	0.579	0.333	0.265	0.039
	上面高	0.898	0.392	0.194	0.039
	鼻高	0.759	0.392	0.310	0.039
	鼻宽	0.282	0.282	0.018	0.040
	眶 宽	0.877	0.411	0.552	0.039
	眶 高	0.802	0.349	0.199	0.039
	面角	0.796	0.425	0.461	0.039
	齿槽点角	0.690	0.395	4.467	0.031
	鼻根点角	0.778	0.377	0.625	0.038
	颅指数	0.918	0.427	0.254	0.039
	颅长高指数	0.511	0.314	0.043	0.040
	颅宽高指数	0.854	0.429	0.353	0.039
	中上面角	0.653	0.358	0.003	0.040
	鼻指数	0.972	0.428	0.866	0.038
	眶指数	0.899	0.414	0.175	0.039
	额宽指数	0.573	0.424	0.236	0.039

由表 15-8 可见,在 18 个未选变量中,齿槽点角的 Wilk's λ 值最小,而且其 F 值为 4.467,大于被选临界值 3.84。因此齿槽点角应作为第 4 个变量被选入模型。前面已提到,当变量齿槽点角进入模型后,其他变量都不符合被选标准,逐步筛选过程结束。

表 15-9 是逐步筛选的总结表,它给出每一步进入或移出的变量的名称,相应的 Wilk's

λ值和F值。可见当颧宽等4个变量被选后，判别函数的Wilk's λ值已降低到0.031，接近零。在表15-9中我们保留了SPSS程序对该表格的注解(英语)。

表15-9　逐步筛选判别分析过程中各步进入或移出的变量以及相应的Wilk's λ值和F值等

Step	Entered	Wilk's λ Statistic	df1	df2	df3	Exact F Statistic	df1	df2	Sig.
1	颧宽	0.221	1	1	19.0	66.874	1	19.0	0.000
2	耳上颅高	0.084	2	1	19.0	98.775	2	18.0	0.000
3	最小额宽	0.040	3	1	19.0	136.369	3	17.0	0.000
4	齿槽点角	0.031	4	1	19.0	124.251	4	16.0	0.000

At each step, the variable that minimizes the overall Wilk's Lambda is entered.

a Maximum number of steps is 42.

b Minimum partial F to enter is 3.84.

c Maximum partial F to remove is 2.71.

d F level, tolerance, or VIN insufficient for further computation.

4. 表15-10显示殷墟颅骨的实例中判别函数的有效性检验，Wilk's λ＝0.031，很小、显著性水平很高，可见两判别组中心的判别得分值差别显著，说明判别函数的有效性高。

表15-10　判别函数的Wilk's λ值和判别函数的有效性检验

Test of Function(s)	Wilk's Lambda	Chi-square	df	Sig.
1	0.031	58.951	4	0.000

5. 表15-11和表15-12分别给出标准化和非标准化的判别函数系数。

表15-11　标准化判别函数系数

	Function 1
颧宽	－1.399
耳上颅高	0.730
最小额宽	1.348
齿槽点角	－0.571

表15-12　非标准化判别函数系数

	Function 1
颧宽	－0.923
耳上颅高	0.535
最小额宽	0.607
齿槽点角	－0.452
(Constant)	5.359

由此非标准化的判别函数可写成

$$F = 5.359 - 0.923(\text{颧宽}) + 0.535(\text{耳上颅高}) + 0.607(\text{最小额宽}) - 0.452(\text{齿槽点角}) \tag{15-18}$$

程序还输出两个先验组组中心的判别得分分别为 5.055 和 -5.560。利用式(15-18)可以计算每个实体的判别得分，无论实体是否进入分析阶段，从而根据它离哪个中心更近以判断它应归入哪组。

对比表 15-12 和全选模型中的结构矩阵(表 15-5)可见，逐步筛选方法所选的颧宽等 4 个变量都是在全选模型的结构矩阵中排列靠前的，而且是与区分北亚和东亚类型颅骨的颧宽度和颅、面高度等特征有关的变量。这表明两种方法的共同性和判别分析结果的稳定性和可信性。

6. 逐步筛选程序输出的"Casewise Statistics"表，显示每一个实体的原始先验分组和判别归组，其判别得分，到两组中心的马氏距离，归属到各组的概率等。值得提出的是无论是未进行验证(Validation)的判别分析和进行"Leave-one-out"的验证，逐步筛选判别的结果都是 21 组颅骨的先验分组和判别归组完全一致，判别正确率都是 100%。这说明对于所分析实例而言逐步选择模型比全选模型的判别有效性更高，尽管逐步选择模型仅选取了颧宽等 4 个变量进入判别函数。顺便指出逐步选择方法也是将高加索组颅骨归入东亚组，与全选模型的判别结果一致。

15.7 多总体判别分析——商周时期原始瓷的产地溯源

本章前面关于判别分析的原理，方法讨论以及实例应用都是局限于两个总体的情况。本节将讨论实体群分别来自 3 个或 3 个以上的总体，即实体的先验分组为 3 组或 3 组以上情况的判别分析。前两节讨论的两总体判别分析的方法基本上能扩展到多总体的情况，只是需要注意：(1)判别函数的数目将不再是一个，而是扩展到 $(t-1)$ 个，t 是先验分组的数目。(2)每个判别函数所能解释的总方差的百分比是不一样的，因此它们在判别归类中的作用不等。对实体的归类也需同时考虑几个判别函数。(3)判别函数的显著性检验将有所不同。(4)归类结果的图形表示方式也不一样。

下面将通过一个实际例子来阐明多总体判别分析的过程。这是关于商周时期原始瓷产地溯源研究的例子。我国在东汉时开始生产瓷器，是最早生产瓷器的国家。但是更早在商代，在江西吴城、湖北荆南寺、河南郑州商城、小双桥和安阳殷墟等商代遗址中发现了原始瓷器的残存，东汉最初的瓷器生产应该是在原始瓷生产技术的基础上发展而成的。本书作者等(1997)曾用中子活化分析方法测量了上述商代遗址出土的原始瓷片的化学元素组成，并根据这些原始瓷片化学组成的相似性和其他考古资料，提出了上述遗址出土的原始瓷，很可能都是吴城及其周遍地区生产的观点。随后本书作者(2003)又用中子活化分析方法测量了浙江黄梅，安徽南陵牯牛山和清阳苍圆塝，以及广东博罗等地发现的商周时期的原始瓷的元素组成(见表 15-28)。这里我们将用判别分析方法，根据原始瓷样品的化学元素组成，对上述 9 个遗址出土的共 86 片原始瓷片作判别归类研究。研究的内容包括两个方面：首先将吴城(20 片)，黄梅(8 片)，牯牛山(6 片)，苍圆塝(8 片)和

博罗(11片)出土的共53片瓷片作为5个先验组，根据它们的元素组成建立判别函数并进行归类，考察归类的正确率有多高。这5个地点的原始瓷片根据考古资料应该是当地生产的。遗址名后的括号显示该遗址被测量研究的原始瓷片数。第二，根据第一步建立的判别函数对荆南寺(8片)，郑州(10片)，小双桥(4片)和殷墟三期以前(11片)的共33片瓷片归类，考察归类结果能否佐证"这四个地点出土的原始瓷为江西吴城及其周遍地区生产"的观点。共86片瓷片的19个元素含量的测量数据在表15-28中列出(因该表篇幅较大，于本章的最后面列出)。判别分析选取了Al, Ba, Ce, Cr, Cs, Eu, Fe, K, La, Mn, Na, Sc, Th和U等共14个元素，其他元素因为有缺失的测量数据而未选为分析变量。下面使用SPSS软件先进行全选模型的判别分析。

15.7.1　全选模型的多总体判别分析

1. 对于86片原始瓷首先需要建立SPSS数据文件，它包含86个实体，每个实体应该有19个变量，但有的实体的某些变量值因未作测量而缺失，故选择14个变量作为分析变量。(1)打开判别分析的主对话窗口，输入分类变量，确定吴城，黄梅，牯牛山，苍圆塝和博罗等5组为先验组。输入所选取的14个元素为分析变量。选择全选分析模式，即全部14个变量同时进入。(2)在"Statistic"窗口选一元方差分析和Box's M。(3)在"Classify"对话窗口中，选择各组的先验概率 $P(G_i)$ 相等。要求计算和输出各实体的判别得分，马氏距离，归属组别的概率等以及汇总表，选择使用组内协方差矩阵作分析，要求输出判别结果的图形显示。

这里对先验概率 $P(G_i)$ 的选择说明如下：本项分析中5个先验组的样本容量相差较大，吴城组有20片瓷片，其他4个遗址的瓷片数在6—11片之间。我们选择各组的 $P(G_i)$ 相等。如果选择 $P(G_i)$ 正比于样本的容量，其后果必然是增大每个实体，包括未进入分析阶段的实体(如荆南寺、郑州等地的瓷片)归属到吴城组的概率。而本实例研究的目的之一是试图佐证荆南寺，郑州等地出土的原始瓷为江西吴城地区生产的观点，从而选择 $P(G_i)$ 正比于样本的容量，扩大实体归属到吴城组的先验概率显然是不合适的。

2. 程序的执行输出的主要内容如下：

(1) 程序首先对输入的实体总数，用于判别分析的实体数目，每个先验分组中的实体数目和分析变量的名称作汇总，并列表输出。因为用户要求对每个变量作一元方差分析，程序输出一元方差分析的结果如表15-13。

表15-13　各总体变量的均值一致性检验

	Wilk's Lambda	F	df1	df2	Sig.
Al%	0.640	6.744	4	48	0.000
Ba%	0.207	46.057	4	48	0.000
CE	0.583	8.594	4	48	0.000
CR	0.366	20.792	4	48	0.000
CS	0.714	4.798	4	48	0.002
EU	0.447	14.866	4	48	0.000
Fe%	0.626	7.168	4	48	0.000
K%	0.616	7.487	4	48	0.000
LA	0.550	9.836	4	48	0.000

续表

	Wilk's Lambda	F	df1	df2	Sig.
MN	0.516	11.278	4	48	0.000
Na%	0.131	79.895	4	48	0.000
SC	0.487	12.640	4	48	0.000
TH	0.382	19.452	4	48	0.000
U	0.345	22.750	4	48	0.000

由表可见,14个元素都没有通过均值一致性检验,至少在2组间均值有显著差别。如果某个变量通过均值一致性检验,它就不具有判别功能,可以考虑将其从分析变量中剔除。

(2) 各组方差一致性的 Box's M 检验不能进行,因为除吴城组外,其他4组的实体数目太少,均低于变量数。

(3) 程序给出4个判别函数的特征值,如表15-14所示。

表15-14 4个判别函数的特征值表

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	16.689	49.6	49.6	0.971
2	11.370	33.8	83.4	0.959
3	4.577	13.6	97.0	0.906
4	1.021	3.0	100.0	0.711

每个判别函数的特征值是实体对该函数判别得分的组间平方和与组内平方和的比值。每个判别函数的特征值被所有特征值之和去除得到的商值,反映该判别函数所能解释的总平方和的百分比,也是度量该判别函数在各判别函数中的"权重"。在与最大的特征值对应的特征向量的方向上,各组的中心间离散最大,与第二大特征值对应的特征向量给出组中心间离散程度次大的方向。关于特征值和特征向量的详细情况将在第十五章讨论主成分分析时介绍。表15-14中各判别函数是根据其特征值的大小的次序列出,第一个判别函数在实体判别归类中起最主要的作用,第二个次之,第四个,即最后一个判别函数在实体判别中的作用已是很不重要的了。表中最后一列是典型相关系数,表征判别得分与分类变量间的相关程度。

(4) 表15-15是程序执行给出的 Wilk's λ 值表,是对各组判别得分均值的一致性检验,从而检验判别函数的有效性。表题下面第一行中的"1 through 4"表示4个判别函数共同的 Wilk's λ 值,它是4个判别函数单独的 Wilk's λ 值的乘积。最后一行表示单独第4个判别函数的 Wilk's λ 值。Wilk's λ 值可以转换成 χ^2 检验,表中4个检验结果的显著性水平值均小于0.01,说明各组判别得分均值的差别,包括对于单独第4个判别函数各组判别得分均值的差别都是显著的。

表15-15 Wilk's Lambda 表

Test of Function(s)	Wilk's Lambda	Chi-square	df	Sig.
1 through 4	0.000	331.940	56	0.000
2 through 4	0.007	209.840	39	0.000
3 through 4	0.089	102.941	24	0.000
4	0.495	29.897	11	0.002

(5) 表 15-16 和表 15-17 是标准化判别函数和非标准化判别函数的系数。因为先验组的数目是 5 组,共有 4 个判别函数。前面(3)中已经说明,对于实体的判别归类,仅需考虑前两个判别函数即可。

表 15-16 标准化判别函数系数表

	Function			
	1	2	3	4
Al%	-1.316	-0.335	0.444	0.617
Ba%	0.596	0.332	-0.870	-0.355
CE	-1.423	-0.778	-0.229	-0.488
CR	-0.501	0.268	0.393	-0.198
CS	0.420	0.775	0.157	-0.009
EU	0.402	0.752	-0.551	0.923
Fe%	0.273	0.199	0.206	0.264
K%	-0.332	-0.435	0.073	0.207
LA	1.090	-0.515	0.139	-0.016
MN	0.185	-0.247	0.259	-0.359
Na%	0.967	-0.335	0.662	0.156
SC	0.861	0.805	0.242	-0.545
TH	0.458	0.291	-0.608	0.650
U	-0.623	-0.289	-0.058	-0.739

表 15-17 非标准化判别函数系数表

	Function			
	1	2	3	4
Al%	-1.015	-0.258	0.342	0.475
Ba%	73.076	40.782	-106.727	-43.596
CE	-0.098	-0.054	-0.016	-0.034
CR	-0.023	0.013	0.018	-0.009
CS	0.088	0.162	0.033	-0.002
EU	1.051	1.965	-1.439	2.412
Fe%	0.459	0.335	0.346	0.443
K%	-1.033	-1.353	0.229	0.644
LA	0.121	-0.057	0.015	-0.002
MN	0.003	-0.004	0.004	-0.006
Na%	5.602	-1.940	3.832	0.902
SC	0.453	0.424	0.128	-0.287
TH	0.101	0.064	-0.135	0.144
U	-0.531	-0.247	-0.049	-0.630
(Constant)	-0.106	0.275	1.005	0.782

利用表 15-16 和表 15-17 中的系数可以计算每个实体的 4 个判别得分值和各组中心

的判别得分值。表 15-18 列出每组组中心的 4 个判别得分值，是组中心在非标准化判别函数空间中的坐标位置。

表 15-18 5 个判别组中心位置的判别得分（根据非标准化判别函数计算）

	Function			
Predicted Group for Analysis 1	1	2	3	4
1	0.139	3.792	0.931	0.200
2	6.085	-0.959	-3.488	0.374
3	0.722	-1.442	0.422	-2.643
4	2.007	-5.071	3.066	0.741
5	-6.531	-1.723	-1.617	0.268

(6) 程序输出的“Casewise Statistics”表，显示每一个实体的原始先验分组和最大可能和次大可能的判别归组，4 个判别函数的得分，离最可能和次可能组中心的马氏距离，归属到最可能和次可能组的概率等。“Casewise Statistics”表所占篇幅甚大，这里不可能予以列出。从“Casewise Statistics”表可知，只有本属于苍圆塝的 # 36 号实体被误判，被归入牯牛山组，总判对率为 $\frac{52}{53}=98.1\%$。该表也显示用“leave-one-out”方法的验证结果，除 # 36 号实体外，还有属于吴城的 # 86 和 # 137 号实体也被误判，分别被归入牯牛山组和博罗组，总判对率为 94.3%。“Casewise Statistics”表也列出未进入分析阶段的实体的归类结果。对于未进入分析阶段的荆南寺，郑州，小双桥和早于殷墟四期的 33 片瓷片的归类结果是，除郑州的 # 66 和 # 69 号实体被判别归入牯牛山组外，其余的 31 片原始瓷片均归入吴城组。因此根据瓷片的元素组成进行的判别分析支持“荆南寺，郑州，小双桥和殷墟四期以前的原始瓷可能是吴城及其周边地区生产”的观点。当然支持并不是证明，判别分析的结果仅表明，在吴城，黄梅，牯牛山，苍圆塝和博罗等 5 个原始瓷产地中，荆南寺等四地的原始瓷在元素组成方面更接近吴城。

(7)判别分析的结果也可以用图形表示，图15-3是53个实体以第一、第二判别函数

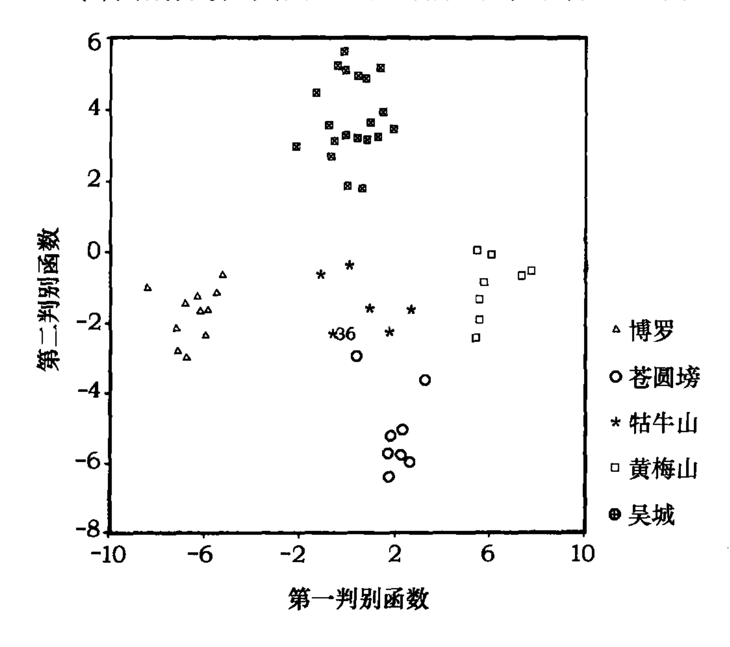


图 15-3 吴城等 5 地 53 片瓷片以第一，第二判别函数为坐标轴的散点图（全选模型）

为轴的散点图。第一和第二判别函数的贡献已占总方差的84.3%。图中显示各组实体间的良好分离,唯一的例外是第四组苍圆塝的#36号实体,它处于第三组牯牛山实体的范围中。

图15-4称为区域图,与图15-3相同也是仅考虑第一和第二判别函数。平面被分成5个区域,对应于5个判别分类。每个实体根据它的第一和第二判别函数的得分决定它处在哪一个区域,即归入哪一组。图上的数字表示组号,每组的中心用"*"符号表示。

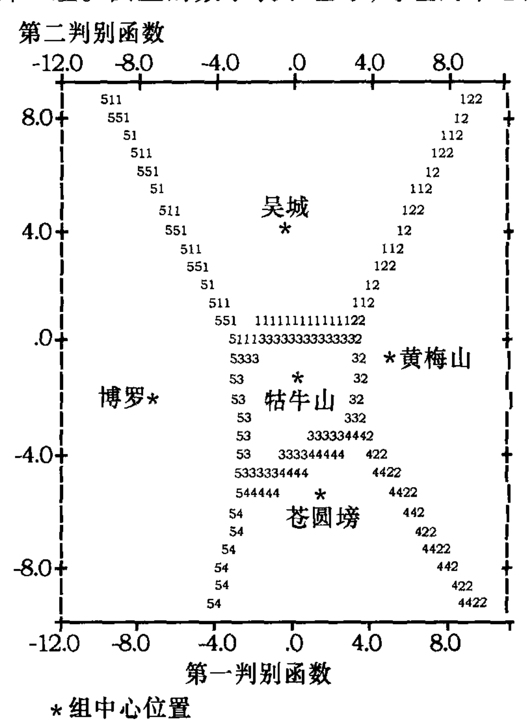


图15-4　吴城等5地53片瓷片多总体判别分析的区域图(全选模型)

(8)原始分析变量对各判别函数贡献的表达,与前述的两总体情况相似,也是通过结构矩阵来显示(表15-19)。结构矩阵记录各判别函数与分析变量间的组内相关系数的计权平均值。

表15-19　结构矩阵

	Function 1	2	3	4
TH	-0.276*	-0.091	-0.226	0.153
U	-0.246*	-0.240	-0.221	-0.048
Na%	0.454	-0.494*	0.303	0.132
CR	-0.057	0.365*	0.148	-0.255
SC	-0.053	0.292*	-0.002	-0.189
K%	0.080	-0.194*	-0.108	0.190
Fe%	0.083	0.181*	0.154	-0.056
CS	-0.099	0.131*	-0.072	0.128

续表

	Function			
	1	2	3	4
Ba%	0.408	-0.103	-0.432*	-0.279
LA	-0.036	-0.101	-0.385*	0.053
EU	0.158	0.116	-0.363*	0.260
CE	-0.050	-0.127	-0.322*	-0.127
MN	0.213	-0.046	0.020	-0.391*
Al%	-0.166	-0.065	-0.047	0.206*

"*"表示每行中绝对值最大的元素,每列带"*"的元素按绝对值大小次序排列

由结构矩阵可见,Th, U, Na 和 Ba 对第一判别函数起主要作用。对第二判别函数贡献大的元素较多,有碱金属 Na, K, Cs,以及 Cr, Sc, Fe 等。3 个稀土元素和 Ba 对第三判别函数贡献较大。而过度元素 Mn,Cr,Al 和稀土 Eu 等对第四判别函数影响较大。

15.7.2 逐步筛选模型的多总体判别分析

多总体情况下逐步筛选变量判别分析方法的讨论仍将通过 15.7.1 节中原始瓷的例子来进行,同样使用 SPSS 程序。因为在 15.6 节两总体逐步筛选判别分析中对于逐步筛选变量的判别过程已作了阐述,因此本节中不予重复,仅对与多总体有关的问题作说明。

1. 分类变量和分析变量的输入同 15.7.1 节的全选模型,但判别分析的主对话框中选逐步筛选方法。与 15.6 节两总体逐步筛选判别分析过程相同,在"Method"对话框中选 SPSS 程序默认的 Wilk's λ 方法,F 标准和 F 临界值。"Classify"对话框中的选项也与前面 15.7.1 节的全选模型的选项一致。

2. 程序执行的输出结果概要说明如下:

(1) 关于各组所属总体的协方差一致性检验。因为对于本项实例,逐步筛选方法最终从 14 个分析变量中仅选择了 7 个,进入模型的分析变量的数目少了,Box's M 检验得以进行。其结果如表 15-20 所示。但是各总体协方差一致性的假设未能被接受。

表 15-20 总体协方差一致性的 Box's M 检验

Box's M		372.982
F	Approx.	2.742
	df1	84
	df2	1979.774
	Sig.	0.000

(2) 变量的筛选过程。根据选定的 Wilk's λ 方法和 F 临界值,变量的筛选过程和相应的 Wilk's λ 值和 F 值总结列于表 15-21。当执行了第七步的筛选,变量 Al 进入模型后,剩下的 7 个变量的 F 值均小于 3.84,程序执行终止。每步筛选后的已进入和未进入变量的容忍度,Wilk's λ 值和 F 值表,鉴于篇幅这里未予列出,它们与表 15-7 和表 15-8 是相似的。

表 15-21　多总体逐步筛选模型过程中变量的进入和移出

Step	Entered	Wilk's Lambda Statistic	df1	df2	df3	Exact F Statistic	df1	df2	Sig.
1	Na%	0.131	1	4	48.0	79.895	4	48.0	0.000
2	Ba%	0.036	2	4	48.0	50.413	8	94.0	0.000
						Approxim F Statistic			
3	U	0.012	3	4	48.0	43.858	12	121.9	0.000
4	SC	0.007	4	4	48.0	35.503	16	138.1	0.000
5	EU	0.004	5	4	48.0	31.381	20	146.8	0.000
6	CE	0.002	6	4	48.0	30.699	24	151.2	0.000
7	Al%	0.001	7	4	48.0	28.422	28	152.8	0.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

(3) 判别函数的特征值和判别分组的均值一致性检验。这分别由表 15-22 和表 15-23 列出。由表 15-23 可见各判别组间实体平均判别得分的差异是显著的。

表 15-22　四个判别函数的特征值和它们的相对贡献

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	11.003	49.5	49.5	0.957
2	7.107	32.0	81.5	0.936
3	3.451	15.5	97.0	0.881
4	0.667	3.0	100.0	0.633

表 15-23　判别分组的均值一致性检验

Test of Function(s)	Wilk's Lambda	Chi-square	df	Sig.
1 through 4	0.001	302.781	28	0.000
2 through 4	0.017	188.463	18	0.000
3 through 4	0.135	92.196	10	0.000
4	0.600	23.514	4	0.000

(4) 标准化和非标准化的判别函数系数。表 15-24 和表 15-25 分别列出标准化和非标准化的判别函数系数,后者有一个常数项。根据表 15-24 或者表 15-25 中的系数建立四个判别函数,并依此可计算每个实体的四个判别得分,其中前 2 个函数对解释总方差的贡献为 81.5%。表 15-26 给出各判别组中心的 4 个非标准化判别函数的得分值。

表 15-24　标准化判别函数系数

	Function			
	1	2	3	4
Al%	-0.603	0.274	-0.486	-0.826
Ba%	0.721	-0.141	0.655	0.408
CE	-0.848	0.892	0.306	0.513
EU	0.504	-0.477	0.499	-1.112

续表

	Function			
Na%	0.555	0.599	-0.614	-0.005
SC	0.550	-0.942	-0.591	0.801
U	-0.280	0.177	0.621	0.260

表 15-25 非标准化判别函数系数

	Function			
	1	2	3	4
Al%	-0.465	0.211	-0.375	-0.637
Ba%	88.507	-17.306	80.395	50.062
CE	-0.059	0.062	0.021	0.036
EU	1.318	-1.246	1.304	-2.906
Na%	3.213	3.469	-3.556	-0.027
SC	0.289	-0.496	-0.311	0.422
U	-0.238	0.151	0.530	0.222
(Constant)	-0.690	-0.699	-1.452	-2.246

表 15-26 各判别组中心的判别得分

	Function			
Predicted Group	1	2	3	4
1	0.446	-2.877	-0.973	-0.158
2	4.673	0.863	3.037	-0.471
3	0.755	0.920	0.201	2.147
4	1.225	4.413	-2.580	-0.438
5	-5.511	0.891	1.328	-0.224

(5) 原始分析变量对判别函数的贡献。SPSS 程序还输出显示反映判别得分与原始分析变量间组内相关系数计权平均值的结构矩阵(表 15-27)。未被选,即未用作分析的变量同样列入表中,但用上标“a” 注明。对 4 个判别函数贡献大的元素分别是(U,Na,Ba),(Na,Sc) ,(Ba,Ce,Eu),(Eu,Al) 与全选模型的结构矩阵的情况接近。

表 15-27 结构矩阵

	Function			
	1	2	3	4
TH[a]	-0.421*	0.193	0.083	0.058
U	-0.326*	0.258	0.270	0.073
CS[a]	-0.264*	0.206	0.245	-0.079
Na%	0.517	0.698*	-0.270	-0.073
SC	-0.038	-0.377*	-0.024	0.196
K%[a]	0.052	0.308*	0.261	-0.014
CR[a]	0.217	-0.281*	-0.008	0.061
Ba%	0.488	0.157	0.544*	0.175
LA[a]	-0.173	-0.063	0.468*	-0.228

续表

	Function			
	1	2	3	4
CE	-0.076	0.135	0.387*	0.094
EU	0.196	-0.139	0.395	-0.462*
Fe%[a]	-0.074	-0.167	-0.050	0.348*
Al%	-0.212	0.060	0.041	-0.231*
MN[a]	0.112	0.008	0.158	0.205*

“*”表示每行中绝对值最大的元素,每列带“*”的元素按绝对值大小次序排列。上标“a”表示该变量未被选中,未进入分析阶段。

(6) 归类结果汇总也由“Casewise Statistics”表列出。显示每个实体的原始先验分组和最大可能和次大可能的判别归组,4个判别函数的得分,离最可能组和次可能组中心的马氏距离,归属到最可能和次可能组的概率等内容。同样因所占篇幅过大,该表未予列出。无论是否进行“Leave-one-out”验证的判别分析,都是2个瓷片被误判。未作验证的判别分析中,吴城的#137瓷片和苍圆塝的#36瓷片均被归类入牯牛山组。“Leave-one-out”验证的归类中,吴城的#137瓷片仍归类入牯牛山组而苍圆塝的#36瓷片被归类入吴城组。两种情况下判别归类正确率均为96.2%。

图15-5是以第一和第二判别函数为坐标轴的53片瓷片的散点图,图中除显示各组的良好分离外,也标出被误判的实体在图上的位置。

对出自荆南寺,郑州,小双桥和殷墟早于四期的33片未参与分析的瓷片的归类结果是,它们全部归类进入吴城组,比全选模型更接近期望结果。图15-6除列出吴城等5地53片参与分析的瓷片外,同时将出自荆南寺、郑州等地的33片未参与分析的瓷片也显示在图中。可以见到这33片瓷片在图中的位置与吴城瓷片的分布区域最相符合。无论是全选模型或逐步筛选变量模型,根据元素组成进行的判别分析都支持“荆南寺,郑州,小双桥和殷墟早于四期的原始瓷可能是吴城及其周边地区生产”的观点。

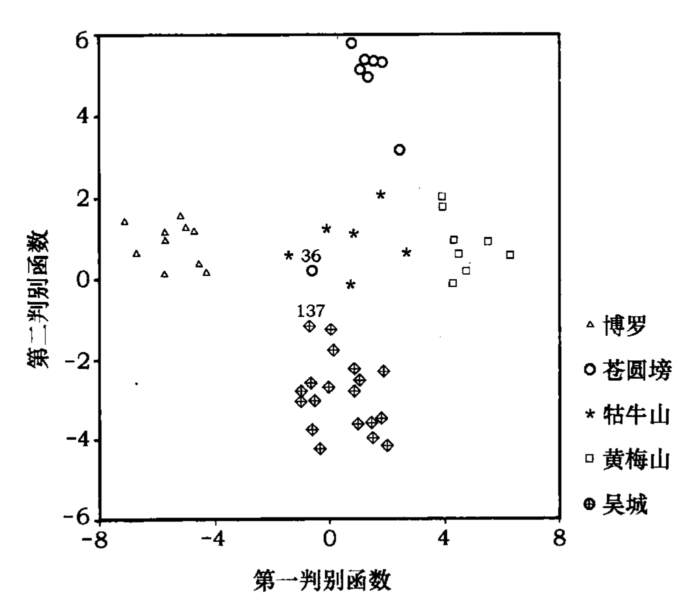


图15-5　吴城等5地53片瓷片以第一,第二判别函数为坐标轴的散点图(逐步筛选模型)

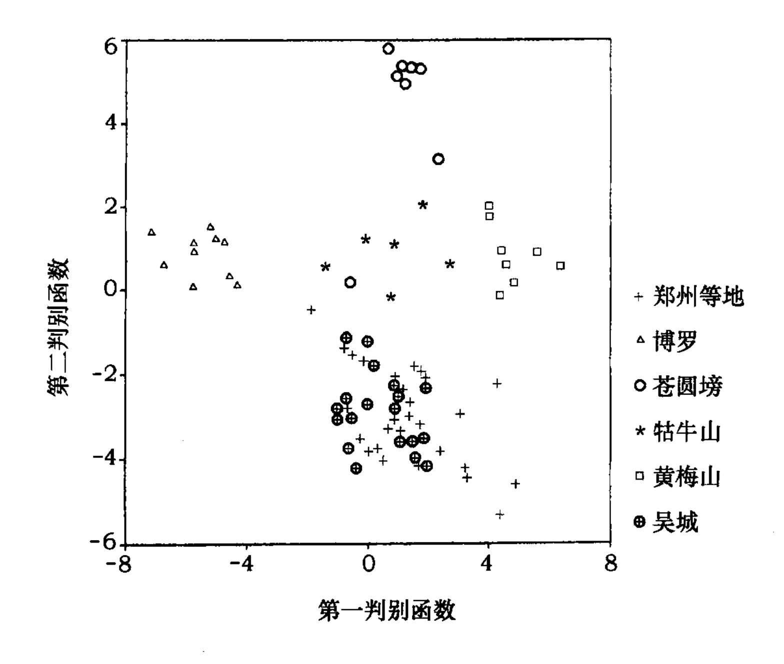


图 15-6 全部 86 片瓷片以第一,第二判别函数为坐标轴的散点图(逐步筛选模型),反映郑州、荆南寺、小双桥和部分殷墟瓷片归入吴城组

15.8 人工神经网络方法应用于实体的归类简介:以我国新石器陶器的归类为例

1999 年 Ma(马清林)等在我国首先利用人工神经网络方法(artificial neural networks,简称 ANN)对我国新石器时代黄河流域,长江流域和南方地区陶器进行归类研究。本节将简要介绍人工神经网络方法应用于实体归类的基本原理和计算方法。

人工神经网络是模拟动物大脑神经网络的结构和行为而发展的一种计算方法。神经网络是由许多个非线性单元组成,称为神经元或节点。神经元具有接受、学习、加工、记忆和传递信息的功能,它们决定了网络的行为。1943 年美国的神经物理学家 McCulloh 和数学家 Pitts 首先提出了神经元机理的数学模型。人工神经网络方法的真正发展是 1982 年 Hopfield 提出离散神经网络模型和稍后的连续神经网络模型,以及 1986 年 Rumelhart 发展了多层次网络和误差反传的网络计算方法以后,它已成为对信息认识、模型建立和预测等应用智能的一门边缘学科,特别适用于非线性的和无明确数学表达式的体系。ANN 已在化学学科中得到广泛的应用,例如模式识别、各种谱图的分析、流程的实时控制、蛋白质结构的预测等。

马清林等用 ANN 于陶器归类的工作属于模式识别。马等工作的原始数据是我国新石器时代黄河流域,长江流域和南方地区三地共 77 片陶片的 9 个主次量元素氧化物的百分含量,他们的研究目标是建立一个神经网络,这个网络通过学习能对未知产地的陶片进行识别,判别它应属于 3 类陶片中的哪一类。

下面通过他们的工作来说明 ANN 中误差反传多层网络的工作原理。这个网络中神经元分为 3 层,分别是输入层、隐蔽层和输出层。因为每片陶片被 9 个元素含量所描述,

输入层相应有 9 个神经元，或节点。马等的文章中未说明隐蔽层所含神经元的数目。一般情况下，如果隐蔽层的节点太少，模型难以正确地传输处理信息，如果节点太多，可能会导致训练过度。在实际工作中可以通过试验来确定隐蔽层节点的数目，为讨论方便，我们假设隐蔽层的节点数为 h。第 3 层是输出层有 p 个节点，在新石器陶器的例子中是 3 个节点，分别代表黄河流域，长江流域和南方三地的陶器。各层的节点间通过“神经”相互联系。

各层神经元之间连接的网络如图 15-7 所示，网络的建立是通过它本身的训练和学习来完成的。网络的学习和信号传递的过程是这样的。(1)将 n 个已知类属的实体的 m 个变量值($x_1, x_2, \cdots x_m$)依次(或一起)输入到输入层的 m 个节点，对于新石器陶器的例子 $m = 9$。(2) m 个信号加以不同的权经“神经”传递给隐蔽层的诸神经元。对于隐蔽层的第 j 个神经元，输入信号是 m 个信号的加权和 $I_j = \sum_{i=1}^{m} x_i m_{ij} + 1 \times b_j$。式中 m_{ij} 是输入层第 i 节点传输信号到隐蔽层的第 j 个节点的权，反映两层的一对神经元之间的联系强度。式中加了一个偏置量 b_j。偏置量 b_j 的引入有利于模型的求解。这个公式也可以看成信号向量 $\boldsymbol{x} = (x_1, x_2, \cdots x_m, 1)$ 和权重向量 $\boldsymbol{w} = (w_{1j}, w_{2j}, \cdots w_{mj}, b_j)$ 间的点积。(3) 隐蔽层的每个节点对输入信号 I_j(如果超过一定的阈)进行“加工”，又把“加工”后的信号加权传递到输出层的各节点。“加工”或转换信号的函数很多，常用的转换函数有 sigmoid 函数：$T_j = \frac{1}{1 + \exp\left(-\frac{I_j}{\theta}\right)}$。这个函数把输出信号限定在 0 与 1 之间，而且能处理非线性模型。(4)输出层的每个节点(新石器陶器的例子中是 3 个节点)接受到 h 个信号后，同样计权加和(也是隐蔽层输出的信号向量和反映隐蔽层节点和输出层的神经元间联系强度的权重向量间的点积)，再用 sigmoid 函数转换后输出。这就是网络输出 $O_k(k = 1, 2, \cdots p)$。(5) 下一步是对网络输出结果与期望值 Q_k' 作比较。期望值可以这样定义：如果输入端输入的是黄河流域的陶片，那么要求黄河流域陶片输出节点的网络输出结果接近 1，而长江流域和南方陶片节点的网络输出结果接近 0。如果输入端输入的是长江流域的陶片，那么要求对应长江流域陶片输出节点的网络输出结果接近 1，而其他 2 个输出节点的输出接近 0。(6) 网络的学习规则经常这样规定，要求实际输出值和期望值之间差值的平方和最小，即 $\sum_{k=1}^{p} (Q_k - O_k')^2 = \min$。所谓学习就是不断地进行迭代计算，调整两组权重向量，即不断改变诸 w_{ij} 和 w_{jk} 值。如果迭代过程收敛，就完成了网络的建立。可以使用相应版本的 MATHLAB 软件执行人工神经网络方法。

马等掌握有三地区的陶片共 77 片，他们从 77 片陶片中选了 49 片的数据作为已知数据输入第一层，用“误差反传学习方法”来训练网络。经过上千次地改变权重值，使得第三层的输出结果与已知的分类结果之间尽量接近。这个过程就是上述的人工神经网络的学习过程。经过训练的网络对剩下的 28 片陶片归类，马等报道归类的符合率达 96%。马等还用该网络对 24 片采自甘肃的新石器时代陶片归类，它们全部被正确地归到黄河流域组。马等也用了主成分分析方法对这批陶片进行了分类，并认为在古陶瓷分类研究中人工神经网络方法更为合适。当然对此可以有不同的看法，但马等的工作是很有意义

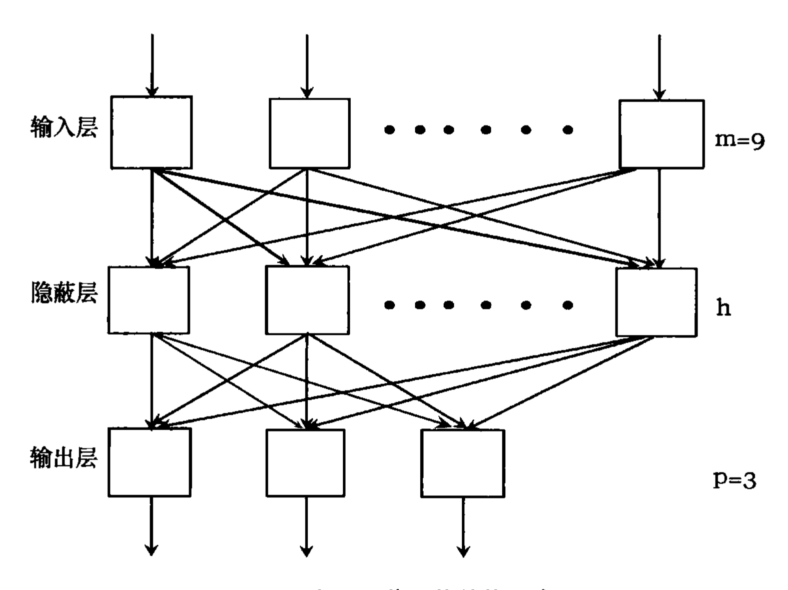


图 15-7 多层反传网络结构示意图

的尝试，是在我国定量考古研究中第一次应用人工神经网络方法。

人工神经网络和判别分析都属于对实体进行归类的多元分析方法。人工神经网络适用于非线性体系和无明确数学表达式的体系，因此其应用面比判别分析广泛。判别分析一般是建立线性的判别函数(原则上也可以用二次多项式或其他形式的函数，但这方面尚研究不够和不易得到通用的软件)，但判别分析的优点是能清楚揭示各原始变量在实体归类中的作用。在考古实体的归类研究中，这两种方法应该是互补的。但是不论使用哪种方法，已知类属的实体的数目均必须足够多，使得模型能受到“良好”的训练。例如为了建立某个神经网络模型，至少需要计算确定($m \times h + h \times k$)个权重的值，显然已知类属的实体的数目太少(原始数据的变量数值为 $n \times m$ 个)，建立的模型将是不稳定，不可靠的。可惜在我国个别应用人工神经网络方法于科技考古资料分析的文章中没有对这个问题予以应有的注意，仅使用很少数量已知类属的实体去训练网络模型，研究结论的可靠性是受到怀疑的。对于人工神经网络感兴趣的读者可阅读罗立强等(1997)的综述，从该文中还可了解到其他有关的文献。

表 15-28 9个地点86片商周原始瓷片的化学组成

瓷片编号	遗址名	先验组	AL %	BA %	CE	CR	CS	EU	FE %	HF	K %	LA	MN	NA %	ND	SB	SC	TB	TH	U	YB
81	吴城	1	8.51	0.037	71.8	111.8	10.32	1.74	1.46	9.2	1.45	44.11	102.3	0.36	43.38	1.47	14.32	0.69	17.29	4.14	3.61
82	吴城	1	10.73	0.051	112.3	92.0	11.23	2.35	3.07	4.6	1.91	55.88	259.0	0.34	38.51	2.52	21.07	0.73	16.62	4.08	3.98
83	吴城	1	7.18	0.042	85.7	125.4	10.80	2.17	1.70	8.6	1.30	50.04	92.6	0.39	50.23	2.28	13.95	0.65	16.86	3.82	3.45
77	吴城	1	9.80	0.038	82.8	96.4	10.41	2.21	2.06	5.3	1.83	55.50	105.1	0.48	37.17	3.29	18.12	0.93	19.56	4.95	3.52
78	吴城	1	7.70	0.027	73.0	93.2	8.11	1.53	1.51	10.1	1.29	48.42	77.4	0.52	43.01	3.13	14.11	1.07	18.62	4.71	3.46
79	吴城	1	10.48	0.035	72.1	106.7	32.36	1.45	1.41	8.6	1.99	47.35	283.1	0.16	35.65	1.81	16.08	0.60	26.49	5.75	4.28
80	吴城	1	9.32	0.041	72.8	100.0	10.15	1.48	2.99	5.5	1.19	39.69	87.2	0.30	39.11	2.84	17.36	0.94	19.31	4.99	2.78
73	吴城	1	9.92	0.034	89.6	118.2	10.20	1.86	3.61	7.1	1.29	53.59	106.3	0.14	50.03	2.18	19.25	0.76	23.65	5.39	3.92
74	吴城	1	7.18	0.040	77.9	90.3	10.39	1.75	1.47	9.1	1.34	46.95	116.0	0.22	31.22	1.74	14.13	0.49	17.25	4.50	3.44
75	吴城	1	7.53	0.044	81.1	85.2	10.79	1.80	1.55	9.3	1.39	44.59	116.3	0.25	33.07	2.94	14.15	0.58	17.49	4.57	3.50
84	吴城	1	11.51	0.056	101.3	95.7	18.08	1.93	2.51	6.5	1.69	59.51	254.9	0.47	41.22	1.70	19.23	0.54	24.62	6.83	3.80
85	吴城	1	8.23	0.041	83.8	167.3	9.63	1.98	1.98	9.3	0.99	44.76	157.0	0.28	32.94	1.77	16.47	0.60	18.58	3.98	3.57
86	吴城	1	11.25	0.052	100.7	155.3	12.80	1.80	2.53	8.1	2.27	60.51	156.1	0.29	49.22	2.44	19.07	0.60	27.46	6.53	4.57
87	吴城	1	8.87	0.030	77.2	93.8	9.38	1.51	1.87	10.1	1.14	38.40	118.4	0.09	37.41	1.14	16.68	0.44	24.25	4.34	3.32
88	吴城	1	11.30	0.053	82.9	51.3	33.33	1.85	2.03	8.2	2.66	47.67	126.0	0.34	44.98	0.95	13.84	0.43	19.68	5.90	3.18
89	吴城	1	7.58	0.028	81.8	93.8	7.78	1.42	1.52	10.0	1.09	45.70	87.8	0.16	40.78	1.66	14.96	0.61	19.70	3.79	3.76
90	吴城	1	7.24	0.031	72.9	75.8	20.91	1.28	4.62	9.2	2.10	38.43	101.4	0.06	32.52	2.98	14.12	0.43	19.26	3.67	3.41
91	吴城	1	8.22	0.032	84.2	89.9	12.97	1.58	1.75	9.8	1.31	51.71	128.5	0.09	38.12	3.35	15.36	0.85	22.19	4.99	3.92
92	吴城	1	10.42	0.049	110.4	180.6	10.36	2.58	2.43	3.4	1.25	65.45	284.0	0.38	71.21	2.18	20.87	1.16	19.24	4.13	4.13
137	吴城	1	10.76	0.062	110.4	73.6	18.41	1.42	3.86	7.5	1.78	54.69	377.9	0.18	40.00	1.70	16.30	0.60	24.74	5.01	4.00
40	黄梅山	2	9.60	0.078	115.7	46.4	11.68	2.94	1.57	7.2	2.27	73.09	184.1	0.80	79.04	1.70	14.42	1.11	20.01	5.34	4.94
41	黄梅山	2	8.65	0.083	109.8	48.0	10.14	2.38	1.79	10.1	2.15	66.51	276.1	0.84	70.45	1.73	13.69	1.06	21.45	6.47	5.02
42	黄梅山	2	8.10	0.090	109.1	45.2	9.28	2.24	1.72	9.7	2.00	67.47	210.1	0.90	68.13	1.70	12.81	1.00	19.09	4.84	4.85
43	黄梅山	2	8.04	0.087	101.3	43.8	9.30	2.74	1.52	11.8	2.20	63.04	311.3	0.97	60.34	1.70	11.90	1.15	19.39	5.33	4.77
44	黄梅山	2	8.61	0.076	111.1	44.0	9.35	2.67	1.60	9.0	2.34	67.22	234.5	0.91	58.48	1.69	12.91	1.00	20.01	5.60	5.00
45	黄梅山	2	8.80	0.074	107.1	43.5	8.90	2.10	1.69	11.9	2.16	64.82	229.6	1.04	50.96	1.63	12.74	1.03	19.22	5.73	4.72
46	黄梅山	2	8.48	0.075	115.1	46.6	8.90	2.38	1.72	11.0	2.04	68.57	219.3	0.98	52.07	2.02	13.02	0.96	19.64	5.84	4.86
47	黄梅山	2	8.69	0.077	118.2	47.6	10.43	3.05	1.72	9.2	2.24	69.38	225.7	0.87	46.10	1.47	14.00	1.47	19.80	5.32	4.72
33	牯牛山	3	8.26	0.070	89.8	64.8	7.67	1.20	1.80	10.1	2.01	49.90	229.0	0.70	44.00	0.83	13.20	1.01	18.00	5.49	3.03
35	牯牛山	3	8.30	0.043	122.0	85.9	12.80	1.39	1.58	8.9	1.48	60.60	156.0	0.50	68.10	0.88	16.90	1.06	22.80	7.62	5.73
36	牯牛山	3	10.10	0.059	111.0	76.7	9.51	1.65	1.97	8.6	1.70	60.10	284.0	0.76	51.20	0.70	15.70	1.00	21.70	7.02	2.89

表 15-28　9个地点86片商周原始瓷片的化学组成(续一)

瓷片编号	遗址名	先验组	AL %	BA %	CE	CR	CS	EU	FE %	HF	K %	LA	MN	NA %	ND	SB	SC	TB	TH	U	YB
37	牯牛山	3	7.65	0.051	93.0	98.1	6.85	1.27	1.99	7.1	1.73	48.60	223.0	0.47	40.90	0.94	13.50	1.00	19.00	4.96	3.39
38	牯牛山	3	8.42	0.063	97.2	72.6	8.00	1.26	1.67	9.3	1.90	50.90	236.0	0.88	46.50	0.72	13.00	0.57	22.00	6.85	2.71
39	牯牛山	3	9.89	0.049	105.0	65.1	7.89	1.48	1.84	10.2	1.66	57.70	270.0	0.71	48.40	0.76	14.90	1.21	20.30	6.47	2.87
25	苍圆塝	4	10.10	0.051	91.0	40.7	8.08	1.39	2.18	8.3	2.10	50.20	209.0	1.31	44.40	0.34	13.40	0.86	15.90	5.39	2.11
26	苍圆塝	4	8.70	0.042	74.7	33.4	6.64	1.14	1.79	8.7	1.76	57.50	211.0	0.40	36.40	0.68	11.00	0.71	13.00	4.80	1.73
27	苍圆塝	4	10.10	0.047	93.7	39.5	7.35	1.30	1.65	8.6	2.12	50.40	171.0	1.41	51.40	0.48	11.50	0.77	21.30	6.65	2.21
28	苍圆塝	4	9.41	0.048	95.3	37.5	6.58	1.30	1.52	9.7	2.34	48.30	230.0	1.50	39.20	0.50	11.40	0.78	23.00	7.05	2.77
29	苍圆塝	4	9.53	0.050	89.9	33.1	5.81	1.25	1.36	3.3	2.05	49.80	168.0	1.42	38.40	0.30	10.20	0.79	21.30	6.82	1.95
30	苍圆塝	4	9.66	0.049	89.7	38.1	6.72	1.19	1.34	2.1	2.12	47.10	194.0	1.41	29.80	0.43	10.20	0.52	19.90	6.13	2.31
31	苍圆塝	4	9.40	0.039	96.8	38.9	7.79	1.30	1.64	2.2	2.40	50.50	193.0	1.51	38.20	0.32	10.90	0.59	22.10	6.77	2.36
32	苍圆塝	4	9.42	0.042	91.0	37.4	7.49	1.22	1.41	3.7	2.07	45.80	186.0	1.41	35.40	0.34	10.70	0.68	24.90	5.80	2.44
138	博罗	5	11.70	0.039	109.0	54.0	14.30	1.20	1.30	13.3	2.10	62.40	99.4	0.09	44.40	0.68	15.00	1.24	41.80	12.00	5.54
139	博罗	5	10.40	0.040	141.0	74.4	18.70	1.81	0.97	10.6	2.17	73.00	64.3	0.10	66.20	0.88	14.80	1.44	26.70	6.15	4.93
140	博罗	5	15.40	0.038	141.0	62.2	15.30	2.70	0.96	14.1	1.80	101.00	79.5	0.07	98.10	0.93	15.70	1.80	44.10	11.20	7.66
141	博罗	5	11.20	0.034	117.0	77.6	15.40	1.42	1.38	13.3	1.84	70.50	117.0	0.09	54.90	1.10	15.10	1.47	39.50	10.70	5.46
142	博罗	5	12.20	0.033	138.0	74.8	16.30	2.32	0.95	11.0	2.12	82.40	89.4	0.09	69.60	0.87	17.80	1.98	40.50	11.20	5.76
143	博罗	5	10.10	0.032	102.0	51.7	16.30	1.27	0.83	13.4	1.97	54.80	78.0	0.10	46.50	0.68	13.70	1.04	36.60	8.35	4.43
189	博罗	5	8.87	0.025	94.2	45.6	13.30	1.06	1.06	14.3	1.83	51.70	89.0	0.10	46.20	0.55	10.50	1.21	21.30	7.75	4.45
190	博罗	5	10.10	0.037	136.0	84.2	14.40	1.94	1.04	12.9	1.68	70.20	34.0	0.08	67.80	0.80	13.90	1.57	26.20	7.40	7.33
192	博罗	5	10.80	0.033	77.6	48.4	13.00	0.84	1.56	14.2	1.60	49.50	105.0	0.07	53.20	0.61	12.00	0.93	25.40	7.41	4.36
193	博罗	5	11.20	0.027	93.6	47.3	12.80	0.96	1.40	14.6	1.66	60.70	88.5	0.08	55.60	0.75	12.30	1.22	30.80	7.30	4.92
194	博罗	5	11.90	0.017	92.4	57.9	10.30	0.94	1.03	14.1	1.31	47.20	63.0	0.06	49.40	1.05	14.00	1.07	39.80	8.90	4.26
48	荆南寺		8.85	0.061	78.4	74.3	10.58	1.59	3.48	n.m.	2.46	50.56	240.0	0.41	n.m.	n.m.	16.12	n.m.	15.74	4.30	3.22
49	荆南寺		8.87	0.058	81.9	67.7	13.10	1.90	2.07	n.m.	1.84	45.90	180.0	0.75	n.m.	n.m.	16.76	n.m.	14.87	4.17	3.67
50	荆南寺		7.72	0.063	78.3	86.0	9.36	1.36	2.52	n.m.	1.77	47.59	180.0	0.22	n.m.	n.m.	13.37	n.m.	15.00	4.70	3.54
51	荆南寺		9.57	0.057	100.6	78.5	17.23	1.95	2.60	n.m.	2.21	53.13	130.0	0.30	n.m.	n.m.	18.80	n.m.	16.88	4.51	4.67
52	荆南寺		9.80	0.057	104.9	79.1	17.14	1.77	2.56	n.m.	2.10	55.03	136.8	0.30	n.m.	n.m.	18.74	n.m.	17.58	3.95	4.57
53	荆南寺		8.36	0.053	102.3	74.0	15.90	1.60	2.54	n.m.	1.70	54.18	120.0	0.30	n.m.	n.m.	18.69	n.m.	17.12	3.70	4.43
54	荆南寺		8.80	0.052	82.0	78.6	10.01	1.37	1.48	10.2	1.54	46.68	95.2	0.41	28.78	n.m.	15.00	0.86	16.03	3.92	3.19
55	荆南寺		10.32	0.072	96.7	86.7	10.01	1.55	2.20	6.3	2.40	53.08	169.0	0.21	25.57	n.m.	19.10	0.79	18.26	4.37	3.91

表 15-28 9个地点86片商周原始瓷片的化学组成(续二)

瓷片编号	遗址名	先验组	AL %	BA %	CE	CR	CS	EU	FE %	HF	K %	LA	MN	NA %	ND	SB	SC	TB	TH	U	YB
103	郑州		8.39	0.036	76.7	82.6	12.72	1.24	2.22	5.0	1.37	43.34	98.0	0.10	0.00	n.m.	14.77	0.81	17.05	4.14	3.41
100	郑州		8.07	0.049	79.5	60.5	10.55	1.35	1.12	7.7	1.72	48.55	92.8	0.16	36.40	1.44	12.99	1.18	19.02	4.55	3.78
101	郑州		6.61	0.031	72.2	64.7	10.86	1.47	1.75	6.3	1.44	45.53	119.5	0.18	37.00	1.43	11.23	0.57	12.87	4.16	3.20
98	郑州		8.01	0.054	77.5	89.5	10.62	1.32	2.66	12.5	1.36	43.78	279.4	0.35	34.40	1.24	14.07	0.80	16.11	4.18	3.24
99	郑州		7.88	0.036	72.0	73.4	9.97	0.86	2.25	10.3	0.98	45.74	259.7	0.23	28.40	1.45	13.03	1.13	15.41	4.04	3.47
102	郑州		8.38	0.053	61.1	70.7	18.60	0.84	1.40	9.3	2.22	38.70	136.0	0.15	27.00	1.73	12.90	1.16	10.90	2.95	2.71
104	郑州		8.02	0.045	83.7	72.8	14.32	1.68	1.10	11.2	1.94	51.23	57.0	0.16	26.41	n.m.	12.49	0.79	21.10	4.39	3.26
105	郑州		8.13	0.031	75.8	62.6	10.74	1.14	1.77	8.4	1.32	47.66	157.3	0.17	21.51	n.m.	10.74	0.53	14.00	3.48	3.20
106	郑州		7.62	0.043	78.4	73.1	18.41	1.24	1.84	10.0	2.01	45.22	139.6	0.28	41.88	n.m.	14.73	1.25	15.89	3.47	3.05
107	郑州		7.28	0.051	76.0	61.9	21.34	1.23	1.49	10.2	2.75	42.66	120.0	0.17	31.80	n.m.	13.32	0.93	13.86	3.19	2.92
94	小双桥		8.69	0.046	57.8	66.7	19.59	1.17	1.80	7.0	1.98	36.74	137.4	0.12	26.80	1.75	12.93	0.77	13.80	3.10	2.65
95	小双桥		9.82	0.049	78.4	85.8	11.83	1.60	2.14	5.8	2.36	44.06	208.8	0.29	25.00	1.83	18.32	0.78	15.99	3.74	3.44
96	小双桥		8.43	0.047	65.1	64.8	19.71	1.48	1.57	7.3	2.18	39.82	121.6	0.23	25.80	1.29	12.81	0.98	13.61	3.09	2.77
97	小双桥		8.25	0.045	102.0	75.7	18.34	2.71	1.87	6.5	1.89	56.96	250.8	0.29	50.40	1.35	14.24	1.64	13.78	4.57	3.84
12	殷墟三期		11.83	0.068	73.7	92.8	22.31	1.94	1.58	5.9	2.83	49.91	85.8	0.47	36.45	4.80	21.30	0.72	19.16	5.73	3.09
13	殷墟三期		11.21	0.046	82.9	93.9	18.24	1.61	1.71	5.7	1.89	50.48	97.6	0.26	41.87	4.18	19.87	0.62	16.99	6.14	3.22
14	殷墟三期		10.34	0.046	78.4	84.6	18.05	1.52	1.51	6.6	2.18	49.14	87.7	0.27	34.53	2.89	17.35	0.70	15.77	4.94	3.20
23	殷墟三期		8.89	0.052	82.0	89.2	19.06	1.80	1.79	7.6	2.10	47.33	88.0	0.21	42.82	2.82	17.55	0.71	17.35	4.73	3.74
7	殷墟三期		11.20	0.065	67.8	85.6	16.63	1.46	2.37	6.3	2.26	44.38	191.1	0.37	28.60	1.45	18.35	0.72	18.71	4.43	2.97
8	殷墟三期		11.72	0.066	84.8	99.3	24.57	1.64	1.53	7.2	2.61	49.13	88.9	0.45	34.30	1.93	21.24	1.24	18.58	4.92	3.26
9	殷墟三期		9.08	0.049	82.8	82.3	17.30	1.52	1.51	6.3	1.92	47.97	77.9	0.22	36.50	1.73	17.04	1.08	17.03	4.64	3.39
10	殷墟三期		11.76	0.044	84.6	92.0	19.98	1.40	1.70	12.2	1.83	50.24	80.5	0.26	32.10	1.66	19.71	1.02	16.10	5.63	2.87
11	殷墟三期		11.62	0.045	91.7	93.8	21.03	1.59	1.83	9.1	1.93	49.61	88.3	0.25	29.20	1.73	20.56	0.47	16.60	5.57	3.54
22	殷墟三期		11.51	0.049	86.1	94.1	20.91	1.35	1.76	10.1	1.78	49.52	85.3	0.24	31.30	1.66	20.08	1.58	16.31	5.45	3.05
24	殷墟三期		10.92	0.082	63.4	86.9	16.44	1.41	2.48	11.4	2.36	42.02	169.8	0.37	28.10	1.42	17.90	1.13	17.91	4.37	2.91

* n.m.表示未测量的缺失值

第十六章　多元数据的降维和主成分分析

本章将介绍应用主成分分析方法于多元数据的降维。当观测数据中有很多个实体，而每个实体又被很多个变量所描述时，直观上是很难从庞大烦琐的数据中观察到其中的现象和规律的。在第十四、十五章曾处理了22组颅骨21个测量指标的数据组和53片商周原始瓷片的14种元素含量的数据组。对于这类复杂的多元数据，不易直观地对实体进行分类和排序。在第十四章中曾利用实体间的相似系数来对实体进行聚类。但是使用相似系数会丢失很多有意义的信息，例如难以了解各个变量在实体分类中所起的作用，也不能分析变量之间或实体之间的相关关系。第十五章在讨论判别分析时，实体的先验分类是已知的，判别分析仅是根据对各实体的观测值来检验实体的先验分类是否符合数据本身的结构，这并不是真正意义上根据实体的属性对它们进行排序和分类。

另一方面在第九章中我们曾见到，当实体仅被2个数值型变量所描述时，实体对于其两个变量的分布情况可以用二维平面上的散点图来表述。散点图能够非常直观地显示出数据的结构，即实体分布的规律。如果各实体点在图上基本上按一条曲线排列，那么这条曲线给出了实体排序的次序，例如在图9-1b上代表一些瓷片的点可以按照瓷片中钾含量的高低来排序。如果各实体点在图上聚集成几个相互分离的集团，那么散点图直接显示实体的分类情况，例如图9-1c显示53片商周时期的原始瓷片可以根据其Ce和Cr的含量分成3组。这是二元变量的情况，当实体被3种属性所描述时，根据实体在三维空间中的分布，依然可以直观地观察到它们分类或排序的规律。但是当属性，或变量的数目多于3个时，就难以再利用散点图来直观地对实体进行分类和排序。

多元数据的降维英语称为Ordination，它是通过某种数学运算找到少数几个（理想情况下是2个或3个）综合变量，并用这些综合变量来描述实体的属性，同时在降维过程中信息量的损失尽可能少。这里所谓信息量是指样本中实体群的总离差，或总方差。我们希望实体群在新的综合变量空间中的离差与它们在原始变量空间中的离差的比值尽可能大，尽可能接近1，或者说新变量能解释的原始总离差的百分比尽可能高。数学上已经建立了多种多样的降维方法，但是其中数学基础最为严格的是主成分分析方法，它也是在各类学科，包括考古学研究中应用最广泛的方法。因此本章只讨论主成分分析一种降维方法。主成分分析英语称为Principal Component Analysis，简称PCA。

我们将在16.2节中较为详细地阐述主成分分析的计算过程。主成分分析的原理及其计算过程涉及矩阵代数，对于不熟悉矩阵运算，而仅为了应用主成分分析方法的读者可以不阅读这一节。在16.1节介绍主成分分析的基本思想时将避开矩阵代数，而只使用矩阵代数的一些术语和符号。

在很多文献中提到另一种重要的降维方法，即因子分析或主因子分析方法。主因子分析和主成分分析有很多共同之处，但也有根本的区别。虽然在考古文献中主要应用主成分分析，本书也不拟讨论因子分析，但在SPSS软件中把主成分分析看成是因子分析中

因子提取的一种方法，即把主成分分析看成是因子分析中的一种方法。而本书要使用SPSS软件进行主成分分析，因此在本章的论述中将会使用因子分析的一些术语，例如主因子提取、因子负载和实体的因子得分等，读者可以理解为主成分提取、主成分负载和实体的主成分得分。在16.4.5节中我们将对这两种方法作比较。本章的最后16.5节将简要介绍对应分析。

16.1　主成分分析的基本思想和分析过程的二维说明

16.1.1　主成分分析的基本思想

我们首先通过一个具体的例子，定性地来阐明主成分分析的基本思想。在20世纪的六七十年代，我国的男性公民基本上都穿同样式样的上衣，称为人民装，国外称之为毛服。上衣的剪裁取决于前身长、后身长、袖长、领宽、肩宽、胸围、腰宽和袖宽等参数，每件上衣这些参数的取值是各不相同的。如果服装厂下料时能够确定几个固定的尺寸进行剪裁，既能适合绝大多数男人的体形，又能提高生产效率和降低生产成本。在测量调查了人们穿着的大量上衣后，会注意到上述的8个参数之间是相关联的，特别是前面3个参数相互之间，以及后面5个参数相互之间是高度相关的，分别可以用“衣长”和“衣宽”两个综合参数来替代。综合参数就是主成分。现在(1)把“衣长”，分为长、中等和短3个衣长尺寸，(2)而每个“衣长”尺寸又分为“肥、正常和瘦”3种“衣宽”型号来剪裁生产9种型号的上衣，那么80%～90%的男子将能买到合身的上衣。也就是说这两个综合参数能够解释总体(全部男子上衣)方差中80%～90%的部分，或者说从8个参数降维到“衣长”和“衣宽”这两个综合参数时，80%～90%的初始信息量被保留了。

当然，这里有一系列的问题要进一步的考虑，(1)根据一批实际测量的上衣参数(样本)，怎样转换为衣长和衣宽两个主成分，即怎样确定变量转换系数(公式(16-3a)中的μ_{kj})。(2)怎样计算每个主成分能解释总方差的百分比，并由此确定选取几个主成分。这涉及每个主成分的特征值的大小以及主成分分析的效率。(3)当确定了被选主成分的数目后(最好只需选二、三个主成分)，需要关心每个原始参数或原始变量的方差中有多少百分比能被解释，即需要了解每个变量的共同度，共同度越高，主成分分析的效率也越高。这涉及计算每个变量对于每个被选主成分的负载量，也就是两者间的相关系数。(4)计算样本中每个实体(上衣)的主成分坐标，称为实体的因子得分，然后分析实体在前几个主成分坐标中的分布规律，从而对实体进行分类或排序。对于上衣剪裁的例子则是根据实体的分布确定上衣的标准尺寸以及缝制每种型号的上衣的比例数等。

设有一个多变量的样本，含有n个实体，每个实体被m个变量$X_j(j=1,2,\cdots m)$所描述，把这组数据写成矩阵形式：

$$X_{(n\times m)}=\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \tag{16-1}$$

矩阵元素x_{ij}是第i个实体第j个变量的取值。每一行数据代表一个实体，也可以看作m维

变量空间中的一个向量；而每一列数据反映一个变量对于 n 个实体的取值，同样可以看作 n 维实体空间中的一个向量。在这里我们设定这些 x_{ij} 都是对变量中心化的，即矩阵每一列的和均为零，即

$$\sum_{i=1}^{n} x_{ij} = 0, (j = 1,2,\cdots m) \tag{16-2}$$

也可以更进一步设定这些 x_{ij} 是对变量标准化或正规化的。原始数据的中心化对于主成分分析的计算是非常重要的。如果数据已经是中心化的，那么矩阵 $\boldsymbol{X}_{(n\times m)}$ 中每列元素的平方和 $\sum_i x_{ij}^2$ 就是变量 x_j 的离差平方和，它等于变量 x_j 的方差 $\mathrm{Var}(x_j)$ 的 $(n-1)$ 倍。而样本中全部变量的总离差平方和是各变量离差平方和的总和，等于即 $\sum_j \sum_i x_{ij}^2$。另一方面变量向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的点积 $\sum_k x_{ki} \cdot x_{kj}$ 是变量 x_i 和 x_j 间协方差 $\mathrm{Cov}(x_i, x_j)$ 的 $(n-1)$ 倍。

现在对 m 维空间的坐标轴作刚性转动，这也等同于对表(16-1)的原始数据作线性转换，即按线性关系另外建立 m 个新变量 y_j：

$$y_{ij} = \sum_{k=1}^{m} x_{ik} \cdot u_{kj}, (i = 1,2,\cdots n), (j = 1,2,\cdots m) \tag{16-3a}$$

式(16-3a)可以写成矩阵相乘的形式：

$$\boldsymbol{Y}_{(n\times m)} = \boldsymbol{X}_{(n\times m)} \boldsymbol{U}_{(m\times n)} \tag{16-3b}$$

$\boldsymbol{U}_{(m\times n)}$ 是一个 $m \times m$ 阶的矩阵，称为变换矩阵。坐标轴的刚性转动，即变量作线性变换时，样本的总离差平方和，或简称总离差是不变的，但是每个变量的离差平方和是在变化的。我们希望在坐标变换后，少数几个新变量 y_j 已能解释样本(16-1)大部分的总离差，而且第一个新变量 y_1 能解释最多的总离差，第二个新变量 y_2 能解释第二多的总离差……，也就是说实体在 y_1 坐标轴方向上的离散度最高，在 y_2 坐标轴方向上的离散度次高……。y_1 和 y_2 分别称为第一和第二主成分，按次序 y_j 称为第 j 个主成分。我们将在 16.2 和 16.3 节中讨论怎样计算得到这些主成分。

16.1.2　主成分分析的二维说明

下面我们将通过一个二维的样本来阐明主成分分析的基本思想和过程，并初步介绍主成分的特征值，变量对于因子的负载，变量的共同度以及实体的因子得分等概念。表 16-1 列出了一个由 8 个实体组成的样本，每个实体被 2 个变量 x_1 和 x_2 所描述。该表的第 2、3 行显示原始变量 x_1 和 x_2 的取值，即 8 个实体的原始二维坐标值，第 4、5 行是 x_1 和 x_2 对变量中心化后的数据。表的第 6、7 行是实体的主成分得分，或称因子得分，后面将说明这 2 行数据是怎样计算得到的。该表的最后一列显示各变量的离差平方和。

表 16.1　二维空间中 8 个实体的例子

实体号	1	2	3	4	5	6	7	8	离差平方和
原始的 x_1	4	4	0	2	-3	-1	3	-5	78
原始的 x_2	6	4	1	0	-4	-4	5	-6	145.5
中心化的 x_1	3.5	3.5	-0.5	1.5	-3.5	-1.5	2.5	-5.5	78

续表

实体号	1	2	3	4	5	6	7	8	离差平方和
中心化的 x_2	5.75	3.75	0.75	-0.25	-4.25	-4.25	4.75	-6.25	145.5
主成分 1	6.71	5.09	0.32	0.67	-5.49	-4.33	5.32	-8.28	217.3
主成分 2	0.51	-0.66	0.84	-1.36	0.36	-1.26	0.74	0.82	6.2

表 16-1 中第 4,5 行的数据是中心化的,因此这二行元素的平方和 $\sum_i x_{ij}^2$ 分别是变量 x_1 和 x_2 的离差平方和,而 $\sum_i (x_{i1}^2 + x_{i2}^2)$ 是样本的总离差平方和,或简称总离差。实际计算得到两个变量的离差平方和分别为 78 和 145.5,它们之间的差别并不大,仅约 2 倍。样本的总离差为 78 + 145.5 = 223.5。

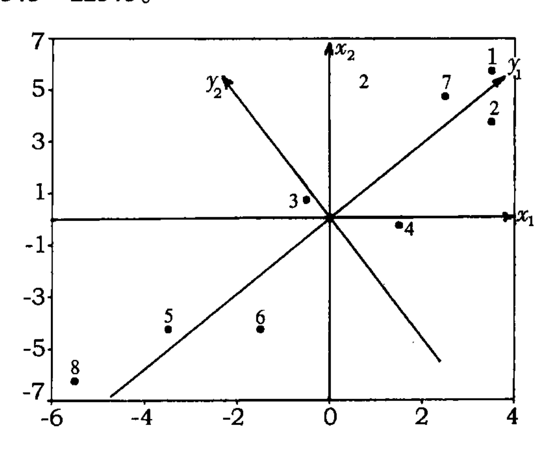


图 16-1　主成分分析概念的二维表示。

图 16-1 是数据中心化后的样本散点图。由图可见,对于这 8 个实体所组成的样本,变量 x_1 与 x_2 间的相关性是较高的,我们将用一个"处于这两个变量之间,又同时与这两个变量高度相关的综合变量"来取代这两个变量,希望这种取代能反映绝大部分的总离差。

现将由 x_1 与 x_2 组成的原始坐标轴刚性转动 θ 角,得到新的坐标系(y_1,y_2)。因为数据已是中心化的,因此坐标轴转动前后每个实体的坐标值(x_1,x_2) 与(y_1,y_2) 间有下面的关系:

$$\begin{aligned} y_{i1} &= x_{i1}\cos\theta + x_{i2}\sin\theta \\ y_{i2} &= x_{i1}(-\sin\theta) + x_{i2}\cos\theta \end{aligned} \quad (i = 1,2\cdots 8). \tag{16-4}$$

根据式(16-4)也可以看到,在新的 (y_1,y_2) 坐标系中样本的总离差是不变的,即有

$$\sum_i y_{i1}^2 + \sum_i y_{i2}^2 = \sum_i x_{i1}^2 + \sum_i x_{i2}^2 \tag{16-5a}$$

因为离差平方和等于方差的 ($n-1$) 倍,所以上式等价于

$$\mathrm{Var}(y_1) + \mathrm{Var}(y_2) = \mathrm{Var}(x_1) + \mathrm{Var}(x_2) \tag{16-5b}$$

在上面的数据转换中,虽然总的离差不变,但是 $\mathrm{Var}(y_1)$ 和 $\mathrm{Var}(y_2)$ 是随 θ 角的变化而变化的。应该这样选择 θ 角,使得 $\mathrm{Var}(y_1)$ 有最大值而 $\mathrm{Var}(y_2)$ 有最小值,即尽量拉开 $\mathrm{Var}(y_1)$ 和 $\mathrm{Var}(y_2)$ 之间的差距。为此需要让 $\sum_i y_{i1}^2$ 或 $\mathrm{Var}(y_1)$ 对 θ 的导数为零。求得的导

数的表达式为

$$\frac{d\sum_i y_{i1}^2}{d\theta} = 2\cos\theta\sin\theta\left(-\sum_i x_{i1}^2 + \sum_i x_{i2}^2\right) + 2\cos2\theta\sum_i x_{i1}\cdot x_{i2} = 0 \qquad (16\text{-}6)$$

将表 16-1 的数据(中心化后的)代入上式,解此方程得到 θ 约等于 54.23 度。即将原始坐标轴转动 54.23 度就得到主成分坐标轴。在图 16.1 中已画出主成分坐标轴 y_1 和 y_2。

得知 $\theta = 54.23$ 度后,可以写出公式(16-3b)中的变换矩阵 $\boldsymbol{U}$。

$$\boldsymbol{U} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} 0.583 & -0.812 \\ 0.812 & 0.583 \end{pmatrix} \qquad (16\text{-}7)$$

变换矩阵又称为因子得分系数矩阵。利用变换矩阵和公式(16-4),可以计算得到各实体在新坐标系的坐标值 (y_{i1}, y_{i2}),称为实体的主成分得分或因子得分。8 个实体的主成分得分已被写入表 16-1 的第 6、7 行。样本的主成分得分也是中心化的,每行元素的平方和分别给出每个主成分的离差平方和,它们分别为 217.3 和 6.2。总离差没有产生变化,依然是 217.3 + 6.2 = 223.5,表明上述的坐标转换不改变总离差的数值。但是两个主成分的离差值的差异明显拉大了,两者差 30 多倍。这样得到的变量 y_1 具有最大的离差及方差 $\mathrm{Var}(y_1)$。y_1 称为第一主成分(PC1),相应变量 y_2 称为第二主成分(PC2)。当我们仅提取第一主成分 y_1,并仅用它来描述样本的各实体时,所保留的信息量的百分比为 $\frac{\mathrm{Var}(y_1)}{[\mathrm{Var}(y_1) + \mathrm{Var}(y_2)]}$。对于上面的例子,第一主成分所保留的信息百分比为 $\frac{217.3}{(217.3 + 6.2)} = 0.972$,从而达到了降维的目的。

在 16.3 节中我们将看到,对应于每个主成分有一个特征值 λ_i,而且 λ_i 是正比于 $\mathrm{Var}(y_i)$ 的。主成分分析中 λ_i 是按数值大小排列的,即有 $\lambda_1 > \lambda_2 > \cdots > \lambda_m$。公式 $\frac{\lambda_i}{\sum_i \lambda_i}$ 给出第 i 个主成分对总离差贡献的百分比。$\frac{(\lambda_1 + \lambda_2)}{\sum_i \lambda_i}$ 和 $\frac{(\lambda_1 + \lambda_2 + \lambda_3)}{\sum_i \lambda_i}$ 是前二个和前三个主成分对总离差贡献的百分比。前二、三个主成分所保留的信息量的百分比是衡量主成分分析有效性的一个重要指标,它取决于原始数据本身的结构。一般情况下,原始变量间的相关性越强,前二、三个主成分所保留的相对信息量也越多。

在主成分分析中我们还希望知道各个原始变量对于所提取的主成分的贡献,称为变量对于主成分(因子)的负载,也就是变量与主成分之间的相关系数。表 16-2 的列出了变量 x_1 和 x_2 对于主成分 y_1,即 PC1 的负载,该表格称为因子负载矩阵表。在 16-2 节中将讨论怎样计算得到这些负载量。

表 16-2 因子负载矩阵表

	y_1 = PC1
x_1	0.973
x_2	0.993

前面计算了第一主成分保留了样本总离差的 97.2%，我们也希望了解所选的第一主成分分别反映单个变量 x_1 和 x_2 的离差的比例，称为变量 x_1 和 x_2 的共同度。可以证明，在提取一个主成分的情况下变量的共同度等于其因子负载的平方。在上面讨论的例子中，变量 x_1 和 x_2 的共同度分别为 $(0.973)^2 = 0.947$ 和 $(0.993)^2 = 0.985$。共同度在 1 和 0 之间变动，它的大小反映了某个变量在所进行主成分分析中作用的大小。

通过阅读上面所述的二维的例子，读者了解了主成分分析的大致计算过程和有关的一些概念，如特征值，因子负载，实体的因子得分，原始变量的共同度等。如果读者不准备更深入地了解主成分分析的计算过程，或者对于矩阵的运算不熟悉，则可以跳过 16.2 节，直接阅读 16.3 节关于主成分分析的应用实例。

16.2　主成分分析的一般计算过程*

本节将更深入地讨论主成分分析的计算过程。一组多元数据如公式(16-1)所示，对其作主成分分析总是从该数据组的方差-协方差矩阵或相关系数矩阵出发的，这两种矩阵都是对称的方阵，因此计算过程涉及对称矩阵的性质和运算，16.2.1 小节将先讨论对称矩阵的性质。

16.2.1　对称矩阵的特征值和特征向量

设有 m 阶方阵 $\boldsymbol{S}_{(m\times m)}$，而且其对称于方阵主对角线的元素相等，即有 $x_{ij} = x_{ij}(i \neq j)$，那么 $\boldsymbol{S}_{(m\times m)}$ 称为对称矩阵。如果 $\boldsymbol{S}_{(m\times m)}$ 的行列式 $|\boldsymbol{S}| \neq 0$，则可以找到 m 个不等于 0 的实数 $\lambda_i \neq 0(i = 1,2\cdots m)$ 和相应 m 个 m 维的列向量 $u_{i(m\times 1)}(i = 1,2\cdots m)$，使得

$$\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{pmatrix}\begin{pmatrix} u_{1i} \\ u_{2i} \\ \vdots \\ u_{mi} \end{pmatrix} = \lambda_i \begin{pmatrix} u_{1i} \\ u_{2i} \\ \vdots \\ u_{mi} \end{pmatrix} \tag{16-8a}$$

λ_i 称为对称方阵的特征根或特征值，而 $\boldsymbol{u}_{i(m\times 1)}$ 是与 λ_i 对应的特征向量。特征值可以按大小排列，使得 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$。上式可以写成

$$\boldsymbol{S}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \quad 或 \ \boldsymbol{S}\boldsymbol{u}_i - \lambda_i \boldsymbol{u}_i = 0 \tag{16-8b}$$

如果将 m 个特征向量作为列向量组成一个 m 阶的方阵 $\boldsymbol{U}$，并将 m 个 λ_i 按大小次序排列组成一个 m 阶的对角线方阵 $\boldsymbol{\Lambda}$，即有

$$\boldsymbol{U} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ u_{21} & u_{22} & \cdots & u_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ u_{m1} & u_{m2} & \cdots & u_{mm} \end{pmatrix} \quad 和 \ \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \lambda_m \end{pmatrix}$$

则公式(16-8a)也可以写成

$$SU = U\Lambda = \begin{pmatrix} \lambda_1 u_{11} & \lambda_2 u_{12} & \cdots & \lambda_m u_{1m} \\ \lambda_1 u_{21} & \lambda_2 u_{22} & \cdots & \lambda_m u_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_1 u_{m1} & \lambda_2 u_{m2} & \cdots & \lambda_m u_{mm} \end{pmatrix} \tag{16-8c}$$

还可以将 m 个特征向量作为行向量组成一个 m 阶的方阵 $\boldsymbol{V}$:

$$\boldsymbol{V} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{m1} \\ u_{12} & u_{22} & \cdots & u_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ u_{1m} & u_{2m} & \cdots & u_{mm} \end{pmatrix}$$

显然 $\boldsymbol{V}$ 是方阵 $\boldsymbol{U}$ 的转置矩阵,即 $\boldsymbol{V} = \boldsymbol{U}^T$。因为 $\boldsymbol{S}$ 和 $\boldsymbol{\Lambda}$ 都是对称矩阵,因此式(16-8c)也可以写作

$$VS = \Lambda V = \begin{pmatrix} \lambda_1 u_{11} & \lambda_1 u_{21} & \cdots & \lambda_1 u_{m1} \\ \lambda_2 u_{12} & \lambda_2 u_{22} & \cdots & \lambda_2 u_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_m u_{1m} & \lambda_m u_{2m} & \cdots & \lambda_m u_{mm} \end{pmatrix} \tag{16-8d}$$

公式(16-8a)至(16-8d)都是等价的。可以证明当 $\boldsymbol{S}$ 是对称矩阵时,由特征向量组成的矩阵 $\boldsymbol{U}$ 是正交矩阵,即有 $\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}^{-1}$,这里 $\boldsymbol{U}^{-1}$是 $\boldsymbol{U}$ 的逆矩阵,即 $\boldsymbol{U}\boldsymbol{U}^{-1} = \boldsymbol{I}$,因此也有 $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{I}$。$\boldsymbol{I}$ 是一个 m 阶的单位方阵,即

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}。$$

对称矩阵的特征值和特征向量有以下的性质

(1) 对称方阵 $\boldsymbol{S}$ 各特征值的平方和等于该方阵各元素的平方和,即

$$\sum_i \lambda_i^2 = \sum_i \sum_j s_{ij}^2 \tag{16-9}$$

(2) 对称方阵 $\boldsymbol{S}$ 各特征值的和等于该方阵对角线元素值的和,方阵对角线各元素值的和称为该方阵的追迹。即有

$$\sum_i \lambda_i = \sum_i s_{ii} \tag{16-10}$$

(3) 特征向量 $\boldsymbol{u}_i$ 之间是正交的,即

$$\boldsymbol{u}_i\boldsymbol{u}_j = 0,\text{如果 } i \neq j;\ \boldsymbol{u}_i\boldsymbol{u}_j = 1,\text{如果 } i = j$$

(4) 由特征向量 $\boldsymbol{u}_i$ 组成的正交矩阵乘任何空间向量$\boldsymbol{x}$,其作用是坐标轴的刚性转动。

(5) 对称方阵 $\boldsymbol{S}$ 的各特征向量组成的正交矩阵 $\boldsymbol{V}$ 和 $\boldsymbol{U}$,均能使 $\boldsymbol{S}$ 转换为对角线矩阵 $\boldsymbol{\Lambda}$,即

$$\boldsymbol{U}^{-1}\boldsymbol{S}\boldsymbol{U} = \boldsymbol{\Lambda} \tag{16-11a}$$

$$\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^{-1} = \boldsymbol{\Lambda} \tag{16-11b}$$

前面讨论了对称矩阵的特征值和特征向量以及它们的性质,下面简单说明怎样计算

得到对称矩阵的特征值和特征向量。公式(16-8b)可改写成

$$(\boldsymbol{S}-\lambda\boldsymbol{I})\boldsymbol{u}=(\boldsymbol{S}-\boldsymbol{\Lambda})\boldsymbol{u}=0 \tag{16-12}$$

因为 $\boldsymbol{u}$ 是非 0 的向量,要求式(16-12)为 0,矩阵 $\boldsymbol{S}-\boldsymbol{\Lambda}$ 的行列式应为 0。即

$$|\boldsymbol{S}-\lambda\boldsymbol{I}|=\begin{vmatrix} s_{11}-\lambda & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22}-\lambda & \cdots & s_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ s_{m1} & s_{m2} & \cdots & s_{mm}-\lambda \end{vmatrix}=0 \tag{16-13}$$

这是 λ 的 m 次多项式方程,也称为矩阵 $\boldsymbol{S}$ 的特征方程。如果 $|\boldsymbol{S}|\neq 0$,解此方程,可以得到 m 个不等于0的实数根 λ_i。将 λ_i 代入式(16-12)可求得相应的特征向量。当然解高次多项式方程和随后的计算特征向量都是大量复杂烦琐的计算,现在均由计算机程序来完成,并且往往得到的是近似值。使用 SPSS 软件进行主成分分析时,程序会自动计算原始数据的协方差矩阵或相关系数矩阵的特征根和特征向量。

16.2.2 主成分分析的一般计算过程

主成分分析是希望将公式(16-1)记录的原始数据 $\boldsymbol{X}_{(n\times m)}=\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$,通过公式(16-3a)规定的线性变换 $y_{ij}=\sum_{k=1}^{m}x_{ik}\cdot u_{kj}$,转换为主成分坐标 $\boldsymbol{Y}_{(n\times m)}$,并使得新变量 y_1 反映样本总离差的最大部分,y_2 反映样本总离差的次大部分,依次类推,y_m 反映样本总离差的最小部分。这个转换写成矩阵形式是公式(16-3b)所示的 $\boldsymbol{Y}_{(n\times m)}=\boldsymbol{X}_{(n\times m)}\boldsymbol{U}_{(m\times m)}$。在 16.1.2 节主成分分析的二维说明中,给出了 $\sum_i y_{i1}^2$ 对 θ 求导数的公式(16-6)。简单的代数运算可以证明式(16-6)是 y_1 和 y_2 协方差 $\mathrm{Cov}(y_1,y_2)$ 的表达式的整数倍。因此要求 $\sum_i y_{i1}^2$ 的导数为零等价于要求 $\mathrm{Cov}(y_1,y_2)=0$,也可以说尽量拉开 $\mathrm{Var}(y_1)$ 和 $\mathrm{Var}(y_2)$ 之间的差距等价于要求 $\mathrm{Cov}(y_1,y_2)=0$。因此对于二维的样本,主成分分析相当与寻找一个变换矩阵 $\boldsymbol{U}$,使得

$$\boldsymbol{U}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{U}=\boldsymbol{Y}^T\boldsymbol{Y}=(m-1)\begin{pmatrix} \mathrm{Var}(y_1) & 0 \\ 0 & \mathrm{Var}(y_2) \end{pmatrix} \tag{16 - 14}$$

式中 $\boldsymbol{X}^T\boldsymbol{X}$ 是原始数据的内积系数矩阵 $\boldsymbol{S}$。如果 $\boldsymbol{X}$ 仅是中心化的,那么 $\boldsymbol{S}$ 是变量的离差矩阵。如果 $\boldsymbol{X}$ 是对离差标准化的,那么 $\boldsymbol{S}$ 是相关系数矩阵。如果 $\boldsymbol{X}$ 是对标准差标准化的,即正规化的,那么 $\boldsymbol{S}$ 是相关系数矩阵的 $(m-1)$ 倍。无论使用哪个 $\boldsymbol{S}$ 矩阵,进行主成分分析的过程是相同的。

上述的对二维情况的讨论可以推广到多维的情况。即多元数据的主成分分析也是对 $\boldsymbol{X}$ 作如下的转换,使得

$$U^TX^TXU = U^TSU = Y^TY = (m-1)\begin{pmatrix} Var(y_1) & 0 & \cdots & 0 \\ 0 & Var(y_2) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & Var(y_m) \end{pmatrix} \quad (16\text{-}14a)$$

另一方面，由 16.2.1 节可知，对于对称矩阵有公式(16-11a) $U^{-1}SU = \Lambda$。因此原始数据内积系数矩阵 S 的诸特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_m$ 依次正比于主成分的方差 $Var(y_1) \geqslant Var(y_2) \geqslant \cdots \geqslant Var(y_m)$。以 S 的特征向量作为列组成的矩阵 U 就是将原始数据 X 转换为主成分坐标 Y 的变换矩阵。主成分分析的计算过程首先就是计算内积系数矩阵 S 的特征值 λ_i 和特征向量 u_i。

计算得到了 λ_i 值后，可以计算仅提取前 k 个主成分时保留的信息量百分比为 $\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{m}\lambda_i}$。并决定提取主成分的数目 k。主成分分析的目的是简化数据以便于直观地观察数据内涵的关系和规律，显然希望提取的主成分的数目不大于 3 个而同时保留的信息量百分比又较大，譬如说大于 60%。能否实现这个希望取决于原始数据本身的结构，关键在于要求原始变量间存在较强的相关性。

确定了提取主成分的数目，譬如提取了 3 个主成分，利用变换矩阵可以计算实体在新坐标系中前 3 个主成分的坐标值，即实体的主成分得分。如公式(16-15)所示。

$$\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & y_{n3} \end{pmatrix} = (X_{ij})_{n\times m}\begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ \vdots & \vdots & \vdots \\ u_{m1} & u_{m2} & u_{m3} \end{pmatrix} \quad (16\text{-}15)$$

在公式(16-15)中，仅需写出了变换矩阵，即因子得分系数矩阵的前 3 列。因为每个主成分都是全部原始变量 x_j 的线性组合，不是单个变量的函数，原则上需要分析每个原始变量分别对每个主成分的贡献。为此我们建立一个新的矩阵，称为因子负载矩阵。

$$L_{n\times m} = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ l_{m1} & l_{m2} & \cdots & l_{mm} \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1}u_{11} & \sqrt{\lambda_2}u_{12} & \cdots & \sqrt{\lambda_m}u_{1m} \\ \sqrt{\lambda_1}u_{21} & \sqrt{\lambda_2}u_{22} & \cdots & \sqrt{\lambda_m}u_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \sqrt{\lambda_1}u_{m1} & \sqrt{\lambda_2}u_{m2} & \cdots & \sqrt{\lambda_m}u_{mm} \end{pmatrix} = U\Lambda^{\frac{1}{2}} \quad (16\text{-}16)$$

因子负载矩阵有一些重要的性质。(1)如果 X 是对离差标准化的，那么矩阵 L 的元素 $\sqrt{\lambda_j}u_{ij}$ 是 X 与 $\Lambda^{-\frac{1}{2}}Y$ 间的相关系数，而 $\Lambda^{-\frac{1}{2}}Y$ 是对离差标准化的 Y。也就是说 L 是原始变量和离差标准化的主成分间的相关系数矩阵。因此 $\sqrt{\lambda_j}u_{ij}$ 反映变量 x_i 对主成分 y_j 的贡献，称为变量 x_i 对第 j 个主成分的负载量。证明如下：X 与 $\Lambda^{-\frac{1}{2}}Y$ 间的相关系数矩阵是 $X^T(Y\Lambda^{-\frac{1}{2}})$，利用公式(16-3b)和(16-8c)可以推导得到

$$X^T(Y\Lambda^{-\frac{1}{2}}) = X^T(XU\Lambda^{-\frac{1}{2}}) = X^TXU\Lambda^{-\frac{1}{2}} = SU\Lambda^{-\frac{1}{2}} = U\Lambda\Lambda^{-\frac{1}{2}} = U\Lambda^{\frac{1}{2}} = L \tag{16-17}$$

(2) 由于 U 是正交矩阵，$\sum_i u_{ij}^2 = 1$。矩阵 L 第 j 列元素的平方和等于 λ_j，即等于该主成分的特征值。表明各变量对第 j 个主成分的负载量的平方和是该主成分的离差平方和。

(3) 也可以对矩阵 L 第 j 行元素的平方求和。第九章的公式(9-8)曾证明相关系数的平方等于回归分析所能解释的总离差的百分比。U 是正交矩阵，因此矩阵 L 每行元素的平方和是等于1的，即主成分坐标系能解释原始数据中每个变量的全部离差。在实际的分析中只选取前几个主成分，如果只选取了3个主成分，那么 $\sum_{j=1}^{3} l_{ij}^2 = \sum_{j=1}^{3} \lambda_j u_{ij}^2$ 反映变量 x_i 对所选前3个主成分所起的作用，也是前3个主成分所能解释变量 x_i 的离差的百分比，也称为变量 x_i 的共同度。

16.3　SPSS 软件主成分分析程序的两个考古应用实例

主成分分析在考古研究中已得到较广泛和多方面应用。据 Baxter(1994)对 1994 年前英语文献的统计，主成分分析应用于考古遗物的化学组成分析的 70 篇，应用于器物形态、人骨与兽骨测量指标分析的 28 篇和应用于器物群比较研究的 33 篇。另外有 23 篇论文，其作者称使用了因子分析于考古研究，但其中相当比例实际上是使用主成分分析方法。在这些论文中具有相当影响的是新考古学学派的创始人 Binford 等(1966)应用主成分分析于法国莫斯特石器研究的文章(文中自称为因子分析，实际上进行的是带主成分轴旋转的主成分分析)。该文挑战法国著名旧石器考古学家博德斯(Bordes)的传统观点，认为不同地点观测到的石器组合的差异并不代表不同人群的文化，而是反映同一种人群在不同地点不同季节从事不同的生产活动。在我国，已发表的主成分分析应用于考古研究的论文中，多数也是应用于古陶瓷根据其化学组成的产地溯源研究。有若干篇论文是根据人颅骨和人牙进行种族分类，应用于器物分类排序的仅见一篇(本章的实例二)。在中文的文献中尚未见到应用主成分分析方法于器物群的比较研究的文章。本节将分别介绍主成分分析应用于瓷器的产地溯源和陶器按其形态分类的两个实例，希望读者能通过实例进一步了解主成分分析的原理，计算方法以及在考古研究中可能发挥的作用。分析过程是使用 SPSS 软件的有关程序来完成的，因此本节也将演示怎样使用 SPSS 软件执行主成分分析。

16.3.1　实例一：商周原始瓷产地的溯源研究

第十五章曾对江西吴城(20 片)、浙江黄梅山(8 片)、安徽牯牛山(6 片)、安徽苍圆塝(8 片)和广东博罗(11 片)出土的 5 组共 53 片原始瓷瓷片作了判别分析。这些瓷片曾用中子活化分析方法测量了 Al, Ba, Ce, Cr, Cs, Eu, Fe, Hf, K, La, Mn, Na, Nd, Sb, Sc, Tb, Th, U 和 Yb 等共 19 种元素的含量(测量数据见表 15-28)。本节将使用 SPSS 软件对这组数据进行主成分分析，希望在降维后的二维或三维主成分坐标空间中分析数据的结

构,观察在降维后的主成分坐标空间中上述5组不同产地的瓷片能否被区直观地区分开。下面将依次按照:(1)数据的输入和各选项的确定,(2)程序执行后输出文件的解读和(3)主成分分析结果的讨论等三部分对分析过程作说明。

(一) 程序对话框中的各个选项。

在执行SPSS主因子分析程序前,除输入原始数据,确定分析变量外,还需要确定一系列选择项目,包括确定选用相关系数矩阵还是协方差矩阵进行分析,确定主因子提取的方法和数目,要求程序输出哪些统计量等。此外还可以要求程序检验原始数据整体和每个变量是否适宜于作主成分分析等。具体过程如下:

1. 建立SPSS的数据文件。虽然在前面两节中一直强调在主成分分析中原始数据应该中心化或标准化,但用户建立数据文件时,不必自己作数据的转换。SPSS的主成分分析程序将自动对原始数据作必要的转换。

2. 打开"Data reduction,→Factor"对话窗口(SPSS软件将主成分分析当作因子分析的一种特殊方法)。首先选择和输入分析变量。对于原始瓷片的例子,由于5组瓷片在以Al和Fe为坐标轴的散点图中分离不明显,而Cs, Mn, Sc的"采样适宜度低",因此这5个元素未被选,它们不参加后面的分析过程。只有其他14个元素被选作为分析变量。关于什么是变量的采样适宜度,本章后面将作说明。

3. 打开"Descriptives"子对话框。可以要求程序输出对一系列统计量的计算结果,其中包括单变量的描述性统计,变量间的相关系数矩阵及其显著性水平,相关系数矩阵的行列式和逆矩阵等。建议在"Descriptives"子对话框中选择下面几个统计量:

(1) "Initial solution"选项。这是程序默认的选择,应该是必选的项目。此项选择要求程序输出各变量初始的共同度,每个主成分的特征值以及所能解释的总离差的百分比。

(2) "KMO and Bartlett's Test of Sphericity"选项。这个选项执行两个检验,检验样本整体上是否适宜于做主成分分析。KMO是Kaiser-Meyer-Olkin采样适宜度的简写,它是样本的全部相关系数的平方和与"全部相关系数的平方和与全部偏相关系数的平方和之和"的比值。KMO值是在0与1间变动,它表征偏相关系数相对于简单相关系数是否很小,或者说表征总离差中有多少比例属于公共离差。一般要求KMO值至少大于0.60,希望能大于0.70,如果KMO太低,例如低于0.5,表明样本不适宜于作主成分分析。Bartlett球形检验是检验相关系数矩阵是否是一个单位矩阵,如果相关系数矩阵接近单位矩阵,说明变量间的相关性很低,样本不适宜于作主成分分析。程序给出Bartlett球形检验的显著性水平。

(3) "Anti image"选项。这个选项是检验单个变量是否适宜于主成分分析。选项的执行将输出反映像协方差矩阵和反映像相关系数矩阵。主要应检查反映像相关系数矩阵对角线上的元素,在SPSS的输出文件中,这些元素用上标"a"标志。如果对角线上某个元素的值小于0.5,表明所对应的变量的采样适宜度低。可以考虑将采样适宜度低的变量从分析变量表中删去,并重新执行主成分分析程序。删除采样适宜度低的变量还将提高整套数据的采样适宜度,即提高KMO值。

4. 打开"Extraction" 子对话框。通过本对话框的各选项,用户确定并要求程序怎样来执行主因子分析,主要是确定提取主因子的方法。

(1) "Method"中选"PCA"。因为我们要进行的是主成分分析,这也是程序的默认选择。如果读者希望进行因子分析,有多种提取因子的方法可供选择,例如最大似然法等。

(2) 第二步要在原始数据的相关系数矩阵和协方差矩阵间作选择,选择使用哪个矩阵进行主成分分析。这是一个重要的选择,两者必选其一。选择不同的矩阵将给出不完全相同的分析结果。我们建议使用相关系数矩阵,因为使用相关系数矩阵使得每个分析变量在分析过程中有大致相等的作用,而且也便于对分析结果的解释。16.4.1 小节将对此进一步讨论。

(3) "Extract"栏,由用户决定选取主成分或因子的数目。一般只选取特征值大于"1"的主成分,这也是程序默认的选择。因为在使用相关系数矩阵进行主成分分析时,主成分特征值的平均值为"1"。当然也可以具体规定提取主成分的数目。

(4) 在"Display"栏中,需要选"unrotated factor solution"。因为我们不准备将主成分轴或因子轴作旋转。"Scree plot"可选可不选,如果作了选择,将输出特征值的碎石图,形象地显示前几个主成分的贡献情况。

(5) 程序规定,在用迭代方法计算相关系数矩阵的特征值和特征向量时,如果计算过程收敛不佳,最高迭代次数不超过 25 次。我们不必去改动。

5. "Rotation"子对话框。SPSS 软件提供对主成分轴或因子轴作旋转处理的程序。旋转可以帮助阐释原始分析变量对主成分或因子的贡献情况,但也会带进新的问题和不确定性。本书将讨论的主成分分析应用实例中,均不作主成分轴旋转,因此在"Method"栏中选择"none"。关于因子轴旋转的问题在 16.4.4 小节中将有较详细的讨论。

在"Rotation"子对话框的"Display"栏中,用户可以选择要求显示变量的因子负载散点图,该图有助于对分析结果的解释。

6. "Factor Scores"子对话框中有两个选项:

(1) 可以要求将实体的因子得分作为新变量写入原始数据文件,这应该是必选的项目。有 3 种计算因子得分的方法可供选择。对于主成分分析,而且主成分轴不作旋转,那么 3 种方法计算得到的实体的因子得分是相等的。如果要作选择,可选"Anderson - rubin"方法,此法给出的因子得分的均值为 0,标准差为 1,而且相互间不相关。

(2) 可以要求显示因子得分系数矩阵。

7. "Option"子对话框的选项包括确定对缺失值的处理方式,以及要求程序输出"变量对所选主成分的负载"表时,变量按负载大小排列,这也有利于对分析结果的解释。

完成对上面 5 个子对话框的选项后,点击主成分分析对话框中的"OK"钮,程序即可执行。

(二) 程序执行的输出文件。

1. 因为在"Descriptives"对话框中,我们选择要求显示"KMO"和"反映像矩阵",程序首先列出检验整套数据和每个变量对于主成分分析的适宜度的结果。对于原始瓷片的例子,输出的检验结果如表 16-3 所示。

表 16-3 KMO and Bartlett 检验

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.778
Bartlett's Test of Sphericity	Approx. Chi-Square	811.501
	df	91
	Sig.	0.000

KMO 样本适宜度度量为 0.778,接近 0.8,而 Bartlett 检验的"相关系数矩阵是单位矩阵"的原假设被否定,说明整套数据是适宜于主成分分析的。

反映像相关系数矩阵表因所占篇幅太大,这里不予列出。由该表可见,除 K 和 Na 的反映像相关系数略小,在 0.54 左右外,其他变量的反映像相关系数均大于 0.6。说明所选各变量的采样适宜度是可以接受的。在本小节(一)2. 讨论的主成分分析选项对话框中,我们未将 Cs,Mn 和 Sc 等 3 个元素作为分析变量输入程序。如将这 3 个变量也作为分析变量输入,程序的执行会揭示它们的反映像相关系数小于 0.5,即它们的采样适宜度是受到怀疑的。

2. 程序接着输出各个变量,即 14 个元素的共同度数值表(表 16-4)。变量的共同度是指被主成分或因子所解释的方差百分比。表中第 2 列显示初始共同度,它是全体主成分所解释的每个变量的方差值。因为本实例中采用主成分分析方法提取因子,即进行的是主成分分析,而且分析是从相关系数矩阵出发的,因此各变量的初始共同度均为 1.000。如果选择了协方差矩阵,则初始共同度为各变量的方差值。表中第 2 列显示所选取的前几个主成分所能解释的每个变量方差值的百分比。后面可以看到,本实例中在要求被选取的主成分的特征值大于 1 的条件下,前 3 个主成分被选。因此表 16-4 第 3 列中所显示的,是前 3 个主成分所解释的每个变量方差的百分比。由表可见对于绝大多数变量,3 个主成分的共同度均大于 0.68,只有元素 Hf 的共同度偏低,为 0.534。说明在原始数据降维到 3 个主成分的情况下,对于变量 Hf,只有约 53% 的方差被反映,Hf 对所进行的主成分分析起的作用不大,可以考虑将 Hf 从分析变量表中剔除,重新执行主成分分析。在本例中我们不作改动。

表 16-4 变量的共同度

	Initial	Extraction
Ba%	1.000	0.793
CE	1.000	0.818
CR	1.000	0.713
EU	1.000	0.885
HF	1.000	0.534
K%	1.000	0.687
LA	1.000	0.921
Na%	1.000	0.832
ND	1.000	0.782
SB	1.000	0.730
TB	1.000	0.763
TH	1.000	0.799

续表

	Initial	Extraction
U	1.000	0.864
YB	1.000	0.864

Extraction Method: Principal Component Analysis.

3. 另一个重要的输出表格是“主成分的特征值和主成分所解释的总方差百分比”表(表 16-5)。前面已经阐明,如果分析过程选择了相关系数矩阵,那么每个主成分的特征值等于该主成分所解释的方差值,而全部特征值的总和等于总方差,也等于变量的数目。表 16-5 的左半边列出全部 14 个主成分的特征值以及每个主成分所解释的总方差的百分比。表中各主成分是按照其特征值的大小排序的,表的第四列显示至某行前,所有主成分所解释的方差的累计百分比。因为在要求特征值大于 1 的条件下只有前 3 个主成分被选,表 16-5 的右半边重复显示了前 3 个主成分的特征值和所解释方差的情况。可以看到当选取 3 个主成分时,能解释 78.5%的总方差;如果选取两个主成分时,能解释 60.5%的总方差。应该说在所讨论的实例中,主成分分析的效果还是比较高的。

表 16-5　特征值和被解释的方差

	Initial	Eigenvalues		Extraction Sums of Squared Loadings		
Component t	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.516	39.400	39.400	5.516	39.400	39.400
2	2.957	21.122	60.522	2.957	21.122	60.522
3	2.512	16.944	78.466	2.512	16.944	78.466
4	0.810	5.783	84.249			
5	0.604	4.316	88.564			
6	0.361	2.576	91.140			
7	0.277	1.980	93.120			
8	0.231	1.653	94.773			
9	0.190	1.356	96.130			
10	0.166	1.189	97.318			
11	0.144	1.032	98.351			
12	9.909E-02	0.708	99.058			
13	7.985E-02	0.570	99.629			
14	5.198E-02	0.371	100.000			

4. 因子负载矩阵(表 16-6)显示各初始变量对所选前 3 个主成分的贡献。表中所示是相应变量和主成分之间的相关系数,称为因子负载。某个变量与某个主成分的相关程度越高,即变量的因子负载越大,表明该变量对相应主成分的贡献越大。因为在“Option”对话框中已提出相应要求,表中的变量,即化学元素是按因子负载的大小排列的,这有助于对分析结果的解释。

对于所分析的 5 组原始瓷片的实例由表 16-6 可见,除 Eu 外的 5 个稀土元素和 2 个放射性元素对第一主成分有主要贡献,碱金属和碱土金属对第二主成分贡献最大,Eu,Sb

和 Cr 的贡献主要反映在第三主成分上。顺便指出 Hf 对第一、第二主成分的贡献相近，Cr 对第三、第二主成分的贡献相近，这种情况有时会使解释分析结果有困难。如有需要，可以通过旋转因子轴来改变变量的因子负载，从而便于解释分析结果，当然因子轴的旋转会带入新的问题。

表 16-6 因子负载矩阵

	Component		
	1	2	3
LA	0.913*	0.168	0.236
TB	0.870*	-7.634E-02	1.703E-02
CE	0.866*	0.225	0.129
YB	0.845*	-0.268	0.281
ND	0.841*	8.774E-02	0.258
U	0.740*	-0.159	-0.540
TH	0.691*	-0.418	-0.383
HF	0.548*	-0.436	-0.211
Na%	-0.188	0.869*	-0.203
Ba%	0.153	0.797*	0.366
K%	0.302	0.722*	-0.273
EU	0.398	0.336	0.784*
SB	-0.268	-0.272	0.764*
CR	-0.225	-0.548	0.602*

因子负载也可以通过散点图来显示，称为因子负载图。因为在本例中提取了 3 个主成分，因子负载图是三维的。在三维的因子负载图中观察 14 个变量的分布不很清晰，这里将三维的因子负载图分解成两个二维的因子负载图，分别以图 16-2 和图 16-3 显示。

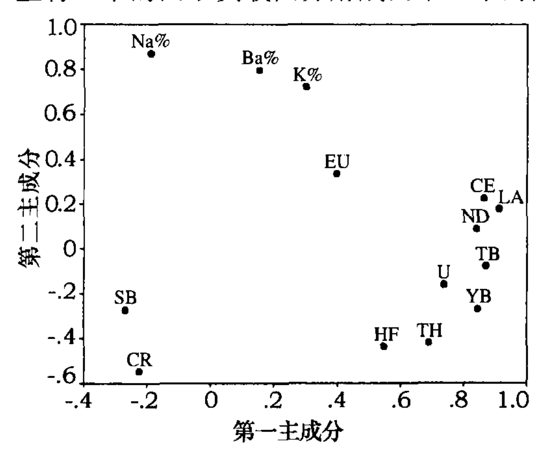


图 16-2 原始分析变量对于第一、二因子的负载图

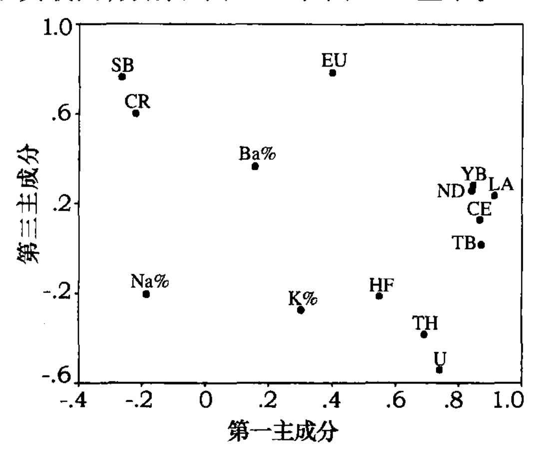


图 16-3 原始分析变量对于第一、三因子的负载图

因子负载图显示了变量间的相关关系以及它们对主成分或主因子的贡献。第一、二因子负载图清楚地显示，对于由 53 片原始瓷片组成的样本，除 Eu 外的 5 个稀土元素以及

U 和 Th 二个放射性元素聚在一起，并在图的右边，说明它们对第一主成分有较大的负载量。这 7 个元素中的多数有相近并接近于 0 的第二主成分负载量，说明它们对第二主成分贡献很小。K，Na 和 Ba 等碱金属和碱土金属聚在图的上部，它们对第二主成分有较大的负载量。Sb 和 Cr 对第一和第二主成分的负载量都是负值，稀土 Eu 介于第一、二组元素集团之间。第一、三因子负载图的解读相对比较复杂，较为明显的现象是除 Eu 外的 5 个稀土元素仍聚在一起，说明这 5 个元素相互间的相关性很强。此外代表 Sb 和 Cr 的点聚得较近，这与它们在第一、二因子负载图的表现相似，Sb，Cr，Eu 以及 U 对第三主成分有较明显的贡献。本节后面还将把因子负载图与实体在主成分坐标系的散点图结合在一起讨论。

5．因子得分系数矩阵和实体的因子得分。

程序输出的因子得分系数矩阵如表 16-7 所示。利用这些系数，根据实体的原始变量值就可以计算每个实体的因子得分，即实体在主成分坐标系中的坐标值，SPSS 软件输出的实体的因子得分是标准化的。在“Scores”对话框中我们已经要求将实体的因子得分作为新变量存入数据文件。程序自动给新变量赋名为“facX-Y”。“X”表示第几个主成分，“Y”表示第几次进行主成分分析的结果。每次执行主成分分析后实体的因子得分作为新变量存入数据文件时，并不改写数据文件中以前已经存入的实体因子得分的数据。

表 16-7　标准化的因子得分系数矩阵

	Component		
	1	2	3
Ba%	0.028	0.270	0.146
CE	0.157	0.076	0.051
CR	-0.041	-0.185	0.240
EU	0.072	0.114	0.312
HF	0.099	-0.147	-0.084
K%	0.055	0.244	-0.109
LA	0.166	0.060	0.094
Na%	-0.034	0.294	-0.081
ND	0.152	0.030	0.103
SB	-0.049	-0.092	0.304
TB	0.158	-0.026	0.007
TH	0.125	-0.141	-0.153
U	0.134	-0.054	-0.215
YB	0.153	-0.090	0.112

6．在分析实体在主成分坐标系的分布前，简单说明 SPSS 主成分分析软件的输出文件中的最后一张表格，主成分（或主因子）的协方差矩阵表（表 16-8）。因为上面的分析过程采用了主成分方法提取因子，即执行了主成分分析，也未作主成分轴的旋转，因此主因子的协方差矩阵就是相关系数矩阵。因为主成分轴是正交的，矩阵对角线上的元素，即主成分的方差总是为 1，而主成分之间的协方差为 0，也是公式（16-14）所要求的。如果用别的方法提取因子，或对因子轴作了旋转，因子的方差和协方差数值将偏离 1 和 0。

表 16-8 主因子协方差矩阵

Component	1	2	3
1	1.000	0.000	0.000
2	0.000	1.000	0.000
3	0.000	0.000	1.000

（三）实体在主成分坐标系中的分布。

主成分分析的重要目的是观察实体在降维后的主成分坐标系中的分布。SPSS 软件的主成分分析输出文件中不包含这类分布图，需要根据程序输出并写入数据文件中的实体的因子得分，即变量 facX-Y 的数值，利用 SPSS 中的作图命令画制相应的分布图。在上面的实例中提取了 3 个主成分。图 16-4 和图 16-5 分别是实体相对于第一、二主成分和第一、三主成分的散点图。

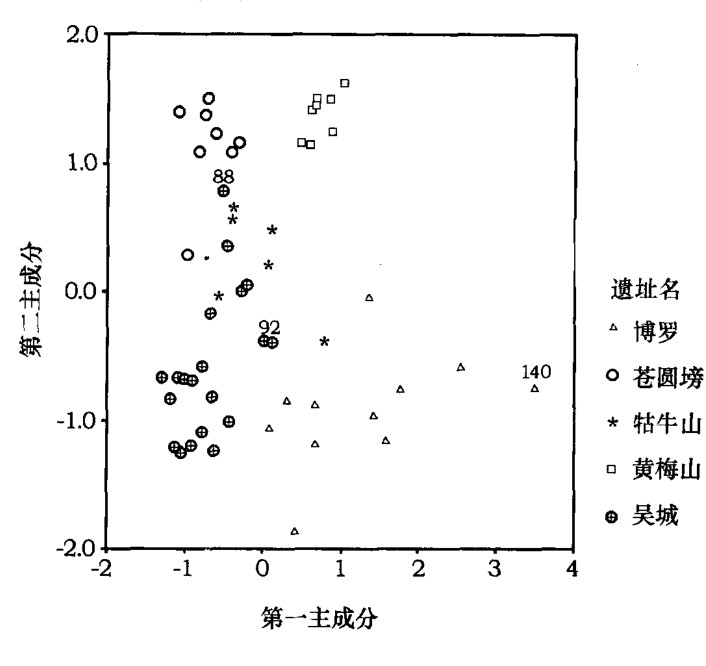


图 16-4 吴城等 5 地 53 片原始瓷片在第一、二主成分坐标中的散点图

实体在两张图中的分布模式非常相似，在图中可以看到 5 组原始瓷片基本上各自聚成组群和各组群间的相互隔离。但是也观察到在各组的边缘部位存在组间实体点的部分混杂和重叠，特别是在第一、二主成分坐标图中牯牛山瓷片点与部分吴城瓷片点的交混较为严重。3 个歧离点在图上已标明。总的情况是各组间的化学元素组成是有明显的差异的，正是这种差异导致了主成分坐标图上各组瓷片聚成基本分离的组群。这种瓷片按照其元素组成分组的结构在原始数据表中是难以察觉的。

主成分分析中另一个需要探讨的问题是各变量（元素）在实体（瓷片）分组中的作用。为此我们对比瓷片在第一、二主成分坐标系的散点图（图 16-4）和第一、二因子负载图（图 16-2）。在图 16-4 中可见博罗和黄梅山瓷片分布在右边，即它们的第一主成分值最大，而图 16-2 显示稀土和铀钍对第一主成分的负载最大，由此可以推论，这些元素在博罗和黄梅山瓷片的含量中的应该是相对偏高的。表 16-9 列出了 5 组原始瓷片中 14 种元素含量

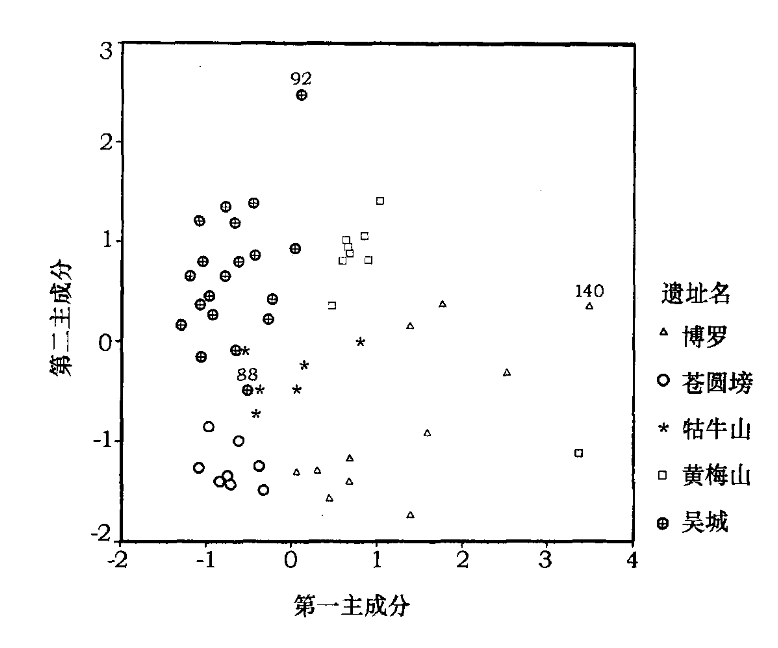


图 16-5 吴城等 5 地 53 片原始瓷片在第一、三主成分坐标中的散点图

的中值,表中元素的排列次序是和因子负载矩阵(表 16-6)中的次序是一致的。由表 16-9 可见,博罗和黄梅山瓷片的稀土和铀钍含量确实比其他 3 组瓷片高。由此我们可以反过来推论:博罗和黄梅山瓷片中稀土和铀钍的高含量导致它们的第一主成分得分高于其他 3 组瓷片,从而导致在第一主成分轴上博罗和黄梅山瓷片与其他地点瓷片的分组。同样的原理可以确定碱金属和碱土金属的含量高使得黄梅山和苍圆塝瓷片有较大的第二主成分得分,这两个地点的瓷片点处于第一、二主成分坐标中的散点图(图 16-4)的上半部分。对比第一、三主成分坐标的实体散点图和第一、三因子负载图揭示,Eu,Sb 和 Cr 的高含量导致吴城和黄梅山的瓷片有较高的第三主成分得分,处于图 16-5 的上半部分。

从前面的讨论中可见,主成分分析不仅能对实体群进行分类,而且能揭示原始变量在实体分类中的作用。主成分分析的这种功能是聚类分析所不能企及的。

表 16-9 5 组原始瓷片中 14 种元素含量的中值$\left(\frac{\mu g}{g}\text{或}\%\right)$

	吴城	黄梅山	牯牛山	苍圆塝	博罗
LA	48.0	67.3	54.3	50.0	62.4
TB	0.61	1.05	1.01	0.74	1.24
CE	83	110	101	91	109
YB	3.59	4.86	2.96	2.26	4.93
ND	39.6	59.4	47.5	38.3	54.9
U	4.64	5.47	6.66	6.39	8.35
TH	19.4	19.7	21.0	21.3	36.6
HF	8.6	9.9	9.1	6.0	13.4
Na%	0.28	0.91	0.71	1.41	0.09
Ba%	0.04	0.08	0.06	0.05	0.03
K%	1.36	2.18	1.72	2.11	1.83

续表

	吴城	黄梅山	牯牛山	苍圆塝	博罗
EU	1.77	2.53	1.33	1.28	1.27
SB	2.18	1.70	0.80	0.39	0.80
CR	94.8	45.8	74.7	37.8	57.9

在实体的散点图中可以观察到存在个别特殊点,它们偏离各自的组中心较远,这些歧离点在图 16-4 和图 16-5 中已被标志。例如在图 16-4 中吴城 88 号实体处于苍圆塝实体的范围,这可能是因为该瓷片的 K 含量偏高,使其第二主成分的得分偏高所致。博罗的 140 号实体的稀土和铀钍含量均偏高,致使该实体点处于图的最右边。吴城的 92 号瓷片也是一个特殊点,该瓷片偏高的 Cr 和 Eu 含量使得其第三主成分值比吴城其他瓷片显著偏大,该实体处于图 16-5 的最上边。怎样处理特殊实体,即是否保留还是剔除这些特殊实体,由研究者决定,我们也将在 16.4.2 中加以讨论。这里仅指出,对于所研究的 5 组瓷片,如果剔除了这 3 个特殊实体,保留的 50 片瓷片在主成分得分散点图的分布,其聚成组群和组群间隔离的情况略优于图 16-4 和图 16-5 所示的结果。需要指出这类特殊实体在聚类分析中是不易被发现的,有时它们可能会导致不适当的聚类结果,这也是为什么作者在实体的分类方法中更偏重于主成分方法的原因之一。

16.3.2 实例二:河南省出土二里岗期前后的陶豆的分期

20 世纪 80 年代作者等(1989)曾尝试应用主成分分析于考古器物的分类研究,具体的对象是河南省出土的 13 件自二里头二期至人民公园期的陶豆。表 16-10 中列出这 13 件陶豆的考古分期和描述其形态特征的变量值,图 16-6 显示了 13 件陶豆的形状。

表 16-10 13 件陶豆的考古分期和描述其形状的测量数据

编号	考古分期	口径/通高	最小径/最大径	通高	盘深/通高	柄高/通高	纹饰
1	二里头二期偏晚	1.06	0.23	21.7	0.26	0.7	1
2	二里头二期偏晚	0.77	0.24	24	0.15	0.81	1
3	二里头四期	0.63	0.32	25.2	0.12	0.88	1
4	二里头二期	0.63	0.28	30.8	0.13	0.9	1
5	晚于二里头四期	1.35	0.52	13.1	0.29	0.65	1
6	同 上	1.75	0.62	10	0.42	0.67	1
7	人民公园期	1.92	0.62	7.8	0.54	0.38	0
8	二里岗期上层	1.8	0.5	9	0.47	0.4	0
9	二里岗期下层	1.35	0.54	12	0.37	0.58	1
a	二里岗期上层	1.33	0.49	12.6	0.52	0.43	0.8
b	早于二里岗期上层	1.14	0.52	15.9	0.29	0.67	1
c	同 上	1.73	0.58	8.5	0.36	0.55	1
d	同 上	1.57	0.68	11.8	0.25	0.45	1

为了使用主成分分析对这批陶豆作分类或分期研究,首先要确定表征陶豆形状的属性,



图 16-6　13 件陶豆的外形图

并对这些属性作定量描述。根据第二章的图 2.2,选择了 6 个属性作定量描述,它们分别为:(1)通高,它反映器物整体大小。(2)口径/通高,这个比值反映陶豆的胖瘦。(3)最小径/最大径,这反映陶豆纵剖面“胖瘦”起伏的程度。(4)盘深/通高,反映豆盘的相对深度和(5)柄高/通高,反映相对柄高。上述 5 个属性都属于数值变量。第 6 个属性反映纹饰,是一个名称属性。为了与前面 5 个数值变量一起作为主成分分析的分析变量,需要将名称变量数值化,为此规定对于有纹的陶豆,该变量取值为 1,无纹的陶豆,该变量取值为 0。13 件陶豆的纹饰多数为弦纹,只有陶豆 # a 为绳纹,对陶豆 # a 该变量取值定为 0.8。13 件陶豆 6 个变量的取值列于表 16-10,它们是进行主成分分析的原始数据。

使用 SPSS 软件主成分分析程序时规定如下的选项:选择相关系数矩阵,要求作 KMO 和变量适宜度检验,采用主成分方法提取因子,主成分轴不作旋转,并要求将因子得分作为变量存入数据文件。分析结果如下:

(1) KMO = 0.873,所有变量的采样适宜度均大于 0.667,说明整套数据和每个变量都适宜于主成分分析。

(2) 第一主成分能解释 78.2%的样本总方差,第二主成分能解释 13.9%的样本总方差,两者一起反映了样本总方差的 92.1%。因此从 6 个原始变量降维为两个主成分变量时,92.1%的信息量被保存。决定选取主成分的数目为 2 个。

(3) 当选取前两个主成分时,程序显示 6 个原始变量的共同度均大于 0.864。说明它们对于所选的两个主成分都有重要的贡献。这从因子负载矩阵也能看出,表 16-11 和图 16-7 是程序输出的变量的因子负载矩阵和因子负载图。表 16-11 每行 2 个元素的平方和正是该行元素的共同度。关于变量的因子负载图,将结合实体在主成分坐标系的散点图进一步讨论。

表 16-11 变量的因子负载矩阵

	Component	
	1	2
口径通高	0.964	0.104
通高	-0.959	-0.220
柄高通高	-0.942	8.072E-02
盘深通高	0.915	-0.165
大小径比	0.838	0.460
纹饰	-0.644	0.729

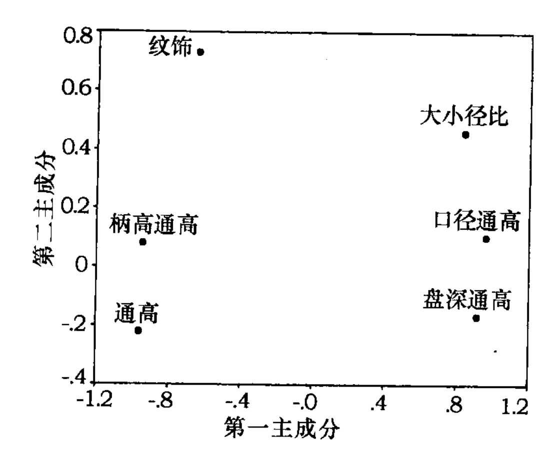


图 16-7 陶豆的主成分分析中原始分析变量的因子负载图

由表 16-11 可见，5 个描述陶豆几何形状的变量对第一主成分有几乎相等的贡献，而纹饰和大小径比对第二主成分的贡献较大。

表 16-12 13 件陶豆的第一和第二主成分得分

陶豆编号	考古分期	第一主成分	第二主成分
1	二里头二期偏晚	-0.80412	-0.63378
2	二里头二期偏晚	-1.26747	-0.5466
3	二里头四期	-1.39335	-0.26069
4	二里头二期	-1.60345	-0.60617
5	晚于二里头四期	0.00419	0.72478
6	同 上	0.54822	1.13511
7	人民公园期	1.55216	-1.40391
8	二里岗期上层	1.20411	-1.80231
9	二里岗期下层	0.248	0.6856
a	二里岗期上层	0.61541	-0.28013
b	早于二里岗期上层	-0.1932	0.57772
c	同 上	0.5878	1.05712
d	同 上	0.5017	1.35326

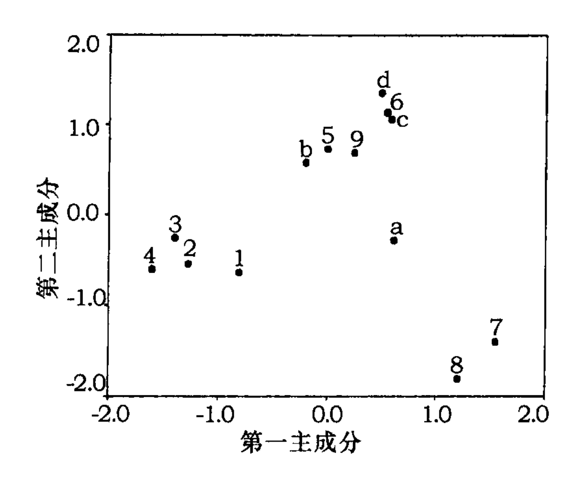


图 16-8 13 件陶豆在第一、二主成分坐标系中的散点图

(4) 根据实体的因子得分(见表 16-12),画出实体在第一、二主成分组成的坐标系中的散点图(图 16-18)。

图 16-8 显示了 13 件陶豆在第一、二主成分组成的坐标系中的分布。由图可见 13 件陶豆分成 3 组,第一组处于图的左边,是 4 件(# 1- # 4)二里头期的陶豆。图的中央偏上的 6 件都是早于二里岗期上层的,组成第二组。而第三组是图的最右边的两件(# 7 和 # 8),它们属于二里岗期上层或更晚的人民公园期。另外 # a 陶豆属于二里岗期上层,在图上处于第二和第三组之间。因此根据所选的描述陶豆的 6 个属性对 13 件陶豆进行主成分分析,得到的陶豆的分组结果与已知的考古分期相符。值得注意的是第一主成分轴反映了陶豆形态随时间变化的过程,也就是说在所研究的例子中,第一主成分轴表现为“时间轴”,为实体按时间排序提供可能。对照图 16-8 和图 16-7,可以看出,从二里头期到人民公园期,陶豆的通高和相对柄高随时间降低,最小直径与最大直径趋于接近,盘深相对变深和口径相对变大的趋势。

上述陶豆的例子是在国内最早使用多元分析方法于考古器物的分类或排序分期研究的尝试。最近滕铭予(2004)较成功地用聚类方法对侯马乔村墓地的几类陶器进行了分期,她同样使用器物线性尺度的比值作为描述器物形状的变量。

16.4 关于主成分分析的几个问题

为了适当地应用主成分分析方法,下面的一些问题应引起注意。

16.4.1 方差-协方差矩阵或相关系数矩阵的选择

执行主成分分析应该使用协方差矩阵还是相关系数矩阵是一个由研究者决定的问题。如果原始数据 $\boldsymbol{X}$ 仅是中心化的,那么其内积系数矩阵 $\boldsymbol{S}$ 是变量的离差矩阵,或方差-协方差矩阵的 $(n-1)$ 倍。如果 $\boldsymbol{X}$ 是对离差或者标准差标准化的,那么 $\boldsymbol{S}$ 将分别是相关系数矩阵或相关系数矩阵的 $(n-1)$ 倍。使用不同的矩阵主成分分析的结果是不完全

一样的。如果使用相关系数矩阵,那么在建立主成分坐标系的过程中所有的变量是等权的,实体的因子得分,即其主成分坐标也是标准化的。如果原始数据仅中心化而没有进一步标准化并使用协方差矩阵,那么各原始变量不再是等权的了,方差大的变量比方差小的变量在分析中起更大的作用。也就是说如果使用协方差矩阵,那么改变原始变量的测量单位会改变主成分分析的结果。使用协方差矩阵时,实体的因子得分是非标准化的。对于应该选用哪一个矩阵,在专门从事考古资料的定量研究的学者之间并没有完全的共识,一般这不是错和对的争论,而是使用哪个矩阵更合理。本书的作者认为 Wright (1989)的建议是值得注意的。Wright 认为,如果实体的属性是测量数据,例如陶瓷的化学组成或者人类或动物骨骼的测量指标,主成分分析时应使用相关系数矩阵,因为不应该无根据地给属性加权。但是如果是计数属性,例如墓葬中各类器物的百分比或频次数,那么可考虑使用协方差矩阵,因为人们不希望给常见的和偶见的器物以相等的权重。此外如果实体的属性包含有多种类型的数据,如测量数据,测量数据的比值,数量化的名称变量等,如 16.3.2 小节的实例二,建议使用相关系数矩阵。SPSS 软件的主成分分析程序默认的选择是使用相关系数矩阵。

16.4.2　歧离实体的处理

如果样本中存在个别偏离样本平均值很大的实体,无论是在个别变量上或者在很多变量上偏离,这些特殊的实体有可能会严重影响主成分分析的结果。因为歧离实体的离差平方值很大,而主成分分析的过程是基于对样本中变量离差的变换。但是随意地将歧离实体从样本中剔除掉,似乎也缺乏理由,有时它们可能反映某种特殊的现象。一种可以考虑的处理方法是,先把特殊实体临时排除在外不参予主成分分析过程,仅用其他实体的数据完成主成分分析并得到的因子得分系数矩阵后,再使用所得到的因子得分系数矩阵来计算这些特殊实体的主成分坐标。然后对全部实体作主成分坐标的散点图,并进行分析。

如果在主成分分析前不易发现歧离实体。可以先对全部实体进行主成分分析,在实体的主成分散点图上可以观察和确定歧离实体。然后剔除这些歧离点,再进行第二次主成分分析,观察二次分析的结果有多大差别。

16.4.3　分析结果的解释

对于分析的最终结果,一般关注两方面的问题。(1)实体在降维后前二、三个主成分坐标系中的分布,对实体进行分类或排序,并作出考古学意义的解释。例如在实例一中,考察比较原始瓷片在主成分坐标系中的分布聚集情况与已知的关于瓷片产地知识间的关系,或者在实例二中考察陶豆按第一主成分轴的排序与陶豆的时代间的关系。(2)探讨原始变量的取值是怎样影响实体在主成分空间中的分类和排序的。为此需要同时分析实体的主成分得分和变量的因子负载。这里可以揭示变量之间的相关关系,揭示相互间相关的变量组在实体分类排序中的作用。因此综合分析实体的分布和变量的因子负载是为了回答,为什么实体的分类或排序会呈现出所观察到的模式,从而有可能揭示更深层次的考古学的现象和规律。这种变量的贡献和实体分布的综合研究正是主成分分

析的一种重要的功能。很遗憾的是,在我国部分已发表的考古和科技考古论文中,主成分分析往往停留在实体的分类和排序阶段,没有进一步去探讨导致出现这种或那种分类结果,其背后隐藏的原因。

另外,有的情况下主成分分析会出现这样一种结果,绝大多数的原始变量对第一主成分都有较显著的贡献,而且都是正贡献。这是因为所有的变量相互间都正相关所致。如果实体是器物,而变量是器物的高度,口径,底径等一些器物线性尺度的测量值,这种情况的出现往往说明第一主成分反映的是器物总体尺寸(Size)的大小。器物的总体尺寸在考古研究中往往并不是最重要的。为了避免这种情况的出现,可以用器物线性尺度测量值的比值来替代直接测量值。在16.4.2小节的实例二中,就是用陶豆的口径与通高的比值,盘深与通高的比值代替口径和盘深等直接测量值。这可以避免出现绝大多数的原始变量对第一主成分都有较显著的正贡献的情况。

16.4.4 主成分轴的转动*

因子轴的旋转是因子分析的一个重要特点,也是它吸引人的优点。旋转能够使得原来在因子轴上负载量大的变量的负载量变得更大,而原来负载量小的变量的负载量更小,从而方便于对分析结果的解释。因子轴作旋转的方法有多种,常用的方法有 Varimax 等。

主成分分析在保留大部分方差的情况下达到了用少数几个主成分来描述实体群的目的。但有时这几个主成分与多个原始变量有较高的相关系数。或者一个变量同时与两个主成分有相近的相关关系,这会引起解释分析结果的困难,即不易解释哪些变量对于所选的哪个主成分作用最显著,从而不易解释哪些变量对于实体的分类或排序有主要的贡献。为了克服这个困难,有的研究者采用对主成分轴作旋转处理的方法。在不改变每个变量的共同度,即不改变每个变量对所选主成分的总负载量的前提下作旋转。

但是对于主成分分析方法应用于考古资料的研究中是否需要旋转主成分轴,文献中存在明显的分歧意见。Baxter(1994)和 Shennan(1997)这两位在考古资料定量研究方面的专家,对于主成分轴旋转均持保留态度,他们的一个重要论据是,旋转将使得各主成分轴失去正交性,而主成分轴的正交性正是主成分分析的重要特点。此外选择不同的旋转方法会得到不完全相同的结果,即主成分轴的旋转将增加对分析结果解释中的主观性和不确定性。Baxter(1994)认为在考古资料的主成分分析中,成功进行主成分轴旋转的例子是个别的。本书讨论的主成分分析应用实例中都不作主成分轴旋转。当然也不能完全排除主成分轴的旋转,有时所选的前几个主成分中的最后两个主成分有相近的特征值,旋转可以拉开它们间的差距,从而确定在它们间剔除哪一个,达到进一步降维的目的。

需要说明主成分轴旋转后,所进行的还是主成分分析,而不是因子分析。相当一些已发表的论文中,包括前面提到的 Binford(1966)所撰写的分析莫斯特石器组合的文章,错误地将主成分轴旋转的主成分分析称为因子分析。在我国,有的应用主成分分析的科技考古研究论文中,也称自己的工作为因子分析。

16.4.5 主成分分析和因子分析*

主成分分析和因子分析(主因子分析)是两种既有联系又有原则差别的降维方法,可

惜在不少文章中把它们混为一谈。为了说明这两种方法的异同,下面先简单介绍因子分析的基本思想。因子分析是 C. Spearman 于 1904 年在美国心理学杂志中首先提出的,后来在心理学研究中得到广泛的应用。譬如随机选择了 n 个学生,进行了代数,几何,物理,化学,生物,地理和地质矿物等 m 个科目测验,现在试图用逻辑推理能力,计算能力,和知识面等 k 个潜在因子来说明每个学生 m 个科目的成绩。潜在因子的数目少于考试的科目($k < m$)。上述的关系可以用公式(16-18)表达。

$$x_{ij} = a_{j1}f_{i1} + \cdots a_{jk}f_{ik} + e_{ij} \tag{16-18}$$

式中 x_{ij} 是第 i 个学生第 j 门科目的成绩,f_{it} 是第 i 个学生第 t 个因子的因子得分。a_{jt} 是第 j 门科目对第 t 个因子的负载量,它与 i 无关。e_{ij} 称为误差因子或每门科目的特殊因子。如果误差因子很小可以忽略不计,那么对每一个学生的评价,主要考察他的推理能力,计算能力,和知识面等少数几个公共因子。因子分析的内容就是要计算得到变量的因子负载量 a_{jt}。公式(16-18)也可以写成矩阵的形式

$$\boldsymbol{X} = \boldsymbol{AF} + \boldsymbol{E} \tag{16-19}$$

可以看出它与主成分分析的基本公式 $\boldsymbol{Y} = \boldsymbol{XU}$(16-14)是不同的。为了计算因子负载矩阵 $\boldsymbol{A}$,需要作一些假设,包括对特殊因子的假设,而主成分分析中计算变换矩阵 $\boldsymbol{U}$ 一般不需要作什么假设。因子分析的理论和方法经历了一系列的发展过程,是不断完善的。如果假设 $k = m$ 和 $\boldsymbol{E} = 0$,那么因子分析与主成分分析等同了。

总的说来两者间有以下的差别:

(1) 主成分分析有严格的数学基础,主成分是原始变量的线性组合;而因子分析中原始变量是潜在的 k 个公共因子的线性组合加上一个特殊因子,其计算过程依赖于某些假设前提,不同的假设会有不同的分析结果。

(2) 原始数据的转换会影响主成分分析的结果,但不影响因子分析。前者寻找方差最大的轴,而后者是寻找相互间协方差大的变量组合。选取主成分的数目不影响前面的主成分,而选取因子的数目会影响前面的因子。

(3) 因子分析的特点是可以作因子轴的旋转来改变分析结果,从而有助于对分析结果的解释。

(4) 如果希望寻找数据中是否存在潜在的,不可直接测量的因素在起作用,建议用因子分析。如果仅希望对多元数据降维,或者对降维后的正交数据用其他统计方法进一步分析,建议使用主成分分析。

16.5 对应分析的简单介绍

对应分析(Correspondence Analysis)是在主成分分析基础上发展起来的一种多元数据的降维方法。

本章前面讨论了对原始数据矩阵 $\boldsymbol{X}_{(n\times m)}$(公式(16-1))进行的主成分分析,称为正分析或 R 分析,得到的 m 个主成分是 m 个原始变量的综合。也可以对原始数据进行逆分析,称为 Q 分析。为此对原始数据矩阵作转置,得到转置矩阵 $\boldsymbol{X}^T_{(m\times n)}$后再进行主成分分析。这时将得到 n 个主成分,它们是 n 个实体的综合。如果每个原始数据值 $x_{ij} \geq 0$,而

且对 $\boldsymbol{X}$ 同时对变量和实体进行了标准化处理，得到 $\boldsymbol{Z}$，那么 $\boldsymbol{Z}\boldsymbol{Z}^T_{(m\times m)}$ 和 $\boldsymbol{Z}^T\boldsymbol{Z}_{(n\times n)}$ 分别是数据标准化后的变量和实体的协方差矩阵，它们的不等于零的特征值的数量相等，而且数值相等。这两个矩阵的特征向量间也有密切的关系。特征值相等的 Q 型和 R 型分析的主成分可以用同一个坐标轴表示，即可以在同一个坐标系中标出实体和变量的散点图，从而看出各类实体的主要特征是什么。这是对应分析的主要优点。在本章前面单纯的 R 分析中，为了了解各类实体的主要特征，即分析各变量对实体分类排序的作用，我们采用的方法是对照实体在主成分坐标系中的散点图和因子负载图。因为这两张图的坐标轴单位是不一样的，不能将它们合并。对应分析给出实体和变量在一起的散点图，便利于同时分析实体的分类排序、变量间的相关和这两者间的关系。但是当实体数和变量数均很多时，同一张图上各类的标记点很多，观察分析都比较费劲，需要对散点图作一些技术性的处理，以便观察分析。

罗宏杰(1997)在古陶瓷化学组成的分析研究中，较多使用对应分析方法，并在他的专著中较详细地介绍了对应分析，有兴趣的读者可以参阅。

第十七章　考古实体的排序和分期

本章将讨论实体的排序和分期问题。排序是将实体依据其某个或某几个属性的取值间相近的程度来排列。实体依据单个变量取值情况的排序是直接明了的，类似于在操场上人们按身材的高矮排队。如果要求综合考虑实体多个变量的取值的相近程度，实体的排序问题就比较复杂，需要用专门的数学方法，例如本章将讨论的 Brainerd-Robinson 方法等。实体的分期是将实体群分成若干组，然后再对组排序，而且排序的标准是按照时间的早晚。考古学研究中涉及的主要是分期问题，包括器物的分期、墓葬和遗址的分期、乃至文化的分期等。在考古学研究中，实体的排序并不是直接的目的。但是如果考古实体，譬如说墓葬，按其在墓地的位置，叠压和打破关系，墓式，出土的器物等特征，已经是排列有序的，同时这个“序”体现了时间的次序。那么通过对有序实体的分划，也是可以实现对它们进行分期的目的。17.2 节将介绍有序实体的最佳分划。本章将分为(1)实体的排序，(2)有序实体的最佳分割和(3)关于渭南史家墓地的分期等 3 部分来论述。

17.1　考古实体的排序

最早从事考古实体排序分期研究的学者之一是著名的英国考古学家皮特里。他提出了顺序年代法，并对埃及前王朝期的法老墓，根据墓中出土的器物进行排序。为此他对每个墓制作一张卡片，上面记录了该墓出土的器物，然后来回排列这些卡片。他的基本出发点是认为每一种器物都经过出现，推广，普及繁荣，衰退和消失等阶段。排列卡片的原则是：(1)使尽可能多的器物，特别是常见的器物，服从上述的演化规律；(2)要求器物从出现到消失所经历的时间尽可能短，即在排列好的卡片序列中所占的区间尽可能短。显然排列这些卡片时，照顾一种器物满足上述规律会影响别的器物的正确排列，因此皮特里的工作是极为费事的，需要多次来回的排列以得到一个“最佳”的序列。皮特里的工作是很经典的，他对法老墓的排序基本上为后来的研究工作所肯定。

20 世纪 50 年代初期，考古学家 Brainerd(1951)和统计学专家 Robinson(1951)合作创造了 B-R 考古实体的排序方法。这是最早使用了数学方法于考古实体的排序研究，这个方法的思路明确，逻辑严格。后又经过 Dempsey(1963)等的改进发展，得到了相当广泛的应用。现通过 Brainerd 和 Robinson 所提出的例子对 B－R 排序方法的原理介绍如下。

17.1.1　Brainerd-Robinson 排序方法的基本原理

假设某个墓地调查了 6 个墓葬(I,II,III,IV,V,和 VI)，其中出土有 5 种形式的器物(A,B,C,D,E)。表 17-1 统计了这 5 种器物在 6 个墓葬中出现的百分比。这是一张器物在墓葬中出现的频率分布表，表中每一行的和均等于 100。

表 17-1　5 种器物在 6 个墓葬中出现的百分比统计，墓葬按原编号排列

墓葬 \ 器物	A	B	C	D	E
I	0	40	0	10	50
II	10	0	50	30	10
III	0	90	0	0	10
IV	60	0	30	10	0
V	0	10	10	60	20
VI	10	20	30	30	10

现要对这 6 个墓葬排序。为此首先要求表中器物的分类是明确的。排序的原则或要求是：(1)希望序列中相邻的墓葬间其器物的百分组成相近，相隔较远的墓葬间器物的百分组成相差也较大；(2)对于所确定的墓葬序列，每种器物经历了发生、发展、极盛和淘汰的正常演化过程。具体的排序过程和对排序结果的检验大致可分为以下三步。

1. 为了实现上述第一个要求，首先要对两个墓葬间器物组成的相近或相异程度作定量的描述，即定义实体间的相似或相异系数(见第十四章 14.3 节)。B-R 方法定义第 i 和第 j 个墓葬间的相似系数 S_{ij} 为

$$S_{ij} = 200 - \sum_{k=1}^{5} | P_{ik} - P_{jk} | \tag{17-1}$$

式中 P_{ik} 是第 k 种器物在第 i 个墓葬中所占的百分数，即表 17-1 主体中第 i 行第 k 列单元格的内容。公式(17-1)求和号后面的每一项定量地反映了某类器物在第 i 和第 j 个墓葬间相异的程度。$| P_{ik} - P_{jk} |$ 值越大，表示 k 类器物在第 i 和 j 墓葬间的百分含量的差值也越大；如果该类器物在两个墓葬中的百分数是相等的，那么这一项为 0。对 5 种器物求和的值是第 i 和 j 个墓葬间总体相异程度的度量，即两座墓葬间的相异系数。如果两个墓葬的器物组成完全一致，求和号后面的每一项均为 0，总和也为 0；如果两个墓葬的器物组成完全相反，即如果某种器物在一个墓葬中出现，它必然在另一个墓葬中缺失，那么这个总和应该等于 200。前面已经说明，每个墓葬中各器物出现频率的和为 100%，即表 17-1 各行的和为 100。因此相异系数是在 0 到 200 间变动，数值愈大，表示两各墓葬的器物组成的差别愈大。公式(17-1)中将“200”被所求得的总和去减，两者之差也是在 0 到 200 间变动，不过现在当两个墓葬的器物组成完全一致时，差值为 200；而当两个墓葬的器物组成完全相反时，差值为 0。因此公式(17-1)给出墓葬之间的相似系数，墓葬间的器物组成愈相似，公式(17-1)给出的数值愈大。因为这种根据器物出现频率表征墓葬间相似程度的方法是 Brainerd 和 Robinson 首先提出的，公式(17-1)计算的结果称为 Brainerd-Robinson 系数(S_{ij})。下面我们计算第 1，2 墓葬间的 B-R 系数，$S_{12} = 200 - (|0-10|+|40-0|+|0-50|+|10-30|+|50-10|) = 200 - 160 = 40$ 依次计算全部 B-R 系数。对于 6 个墓葬，应该有 $(6+5+4+3+2+1) = 21$ 个 B-R 系数。将它们写入下面的表格中，得到墓葬未经重排的 B-R 系数 S_{ij} 表。

这实际上是一个 6×6 的对称矩阵，主对角线两边的元素是相等的，即有 $S_{ij} = S_{ji}$，因此只需写出左下三角的元素即可。

表 17-2 实体间的 *B-R* 系数 S_{ij} 表，实体按原编号排列

	I	II	III	IV	V	VI
I	200					
II	40	200				
III	100	20	200			
IV	20	100	0	200		
V	80	100	40	40	200	
VI	80	160	60	100	120	200

2. 第二步是重新排列墓葬的次序，使得器物组成相近的墓葬，即 B-R 系数大的墓葬相邻排列。这需要规定正确排列的标准。正确的排列应该使得数值大的 B-R 系数靠近主对角线，而数值小的 B-R 系数在左下角。这是因为 $i-j=1$ 的 S_{ij} 是直接相邻两墓葬的 B-R 系数，它们应该数值大，$i-j=2$ 的 S_{ij} 是中间有一墓葬相隔的两墓葬的 B-R 系数，它们应该次大。而 S_{16} 是正确排列的首尾两个墓葬的 B-R 系数，它应该最小。用严格的数学语言来表述上面的思想是要求：

(1)表 17-2 中每一条平行于主对角线的斜线上各 S_{ij} 的平均值 M_i 按照离主对角线的远近应该单调下降。其中

$$M_i = \frac{1}{n-i+1}[S_{i1} + S_{(i+1)2} + \cdots + S_{n(n-i+1)}],(i = 1 \to n, i = 1 \text{ 对应主对角线}) \tag{17-2}$$

单调下降是要求 $M_i \geqslant M_{i+1},(i = 1 \to n)$ (17-3)

(2) 这些 M_i 的和最小，即

$$D = \sum_{i=1}^{n} M_i = \min \tag{17-4}$$

表 17-2 中的墓葬是按照原编号排列的，可以计算得到 $M_1 \to M_6$ 相应为 200,44,85,60,120,80，显然不满足公式(17-3)的要求。后面可见到这 6 个 M_i 值之和 $D = 589$ 也不满足公式(17-4)。因此墓葬需要重新排列，使得公式(17-3)和(17-4)的要求得到满足。这里的计算工作量是很大的，在所分析的例子中有 6 个墓葬，应该有 $\frac{6!}{2} = \frac{(6\times5\times4\times3\times2\times1)}{2} = 320$ 种不同的排列方法。如果墓葬数更多，可能的排列方法将按阶乘函数($n!$)增加。因此只能由计算机来寻找满足 B-R 准则的排列次序。

根据公式(17-3)和(17-4)规定的原则，对表 17-2 的 6 个墓葬进行重新排列得到表 17-3 所示的排列次序。

可以看出在表 17-3 中数值大的 B-R 系数均处于主对角线的邻近，表的左下部位集中了数值小的 B-R 系数。可以计算 $M_1 \to M_6$ 相应为 200,112,80,47,20 和 0，满足公式(17-3)的要求。这 6 个 M_i 的和 $D = 459$，明显小于表 17-2 的 $D = 589$。可以证明表 17-3 的 $D = 459$ 是满足式(17-4)的。

表 17-3　实体按照公式(17-3)和(17-4)要求墓葬重新排列后的 B-R 系数表

	III	I	V	VI	II	IV
III	200					
I	100	200				
V	40	80	200			
VI	60	80	120	200		
II	20	40	100	160	200	
IV	0	20	40	100	100	200

3. 第三步是检验,对于表 17-3 所列出的墓葬排序,5 种器物是否体现了从出现、推广、普及繁荣、衰退和消失的发展规律。为此建立表 17-4,观察墓葬按照 B－R 准则要求排列时 5 种器物的频率分布

表 17-4　5 种器物在 6 个墓葬中出现的百分比统计,墓葬按 B－R 准则要求排列

	III	I	V	VI	II	IV
A	0	0	0	10	10	60
B	90	40	10	20	0	0
C	0	0	10	30	50	30
D	0	10	60	30	30	10
E	10	50	20	10	10	0

为了观察的方便,表 17-4 与表 17-1 相比,行与列作了转置。表 17-4 中每一行记录了每种器物在 6 个按 B-R 准则排列的墓葬中出现的频率,由频率的变化可见,每种器物经历了从出现、发展到消失的过程。因此 B-R 排序的第二个要求也是被满足的。

前面的讨论虽给出了墓葬的 B-R 排列次序,却不能确定排列次序的哪一端时代早,哪一端时代晚。需要有另外的证据来建立墓葬的排列次序与时代早晚间的对应关系,譬如说一对墓葬的叠压关系等。一般情况下,这种对应关系是不难建立的。这里为讨论的方便,假设墓葬 III 是最早的,相应墓葬 IV 就应该是最晚的了。这样从表 17-4 就可以认为"A"是晚期器物,它在墓葬序列的中期才出现。器物"B"在墓葬序列的早期已处于繁荣阶段,在墓葬序列的中晚期衰退消失,它应是早期的器物。器物"D"和"E"在所讨论的墓葬序列中基本经历了从出现到消失的全过程。

Brainerd-Robinson 方法根据考古实体的某些属性出现的频率对实体进行排序,的确提供了一种客观,而且定量的方法。它得到了西方众多考古学家的认同、使用和进一步的发展。B-R 方法和其改进方案至今还得到应用。但使用 B-R 方法,也需要注意:(1)因为方法基于"频率",属性出现的频次数必须足够大。例如在上面讨论的例子中,每种器物在墓葬中的数量要足够多,这样"频率"才稳定和具有统计学的意义;(2)在上面讨论的例子中,认为墓葬的 B-R 系列与时间过程相对应。这要求这些墓葬属于同一或近邻地区时才成立,这样墓葬中器物组成的变化仅取决于时间因素。

在我国公开发表的应用 B-R 方法于考古实体排序尝试的工作有:(1)本书作者(1983)对华北几个晚更新世动物群的排序和裴安平等(1991)对江陵雨台山 34 座日用器组合齐全的和 127 座仿铜礼器组合齐全的墓葬进行的年代序列分析。下面简要介绍这二项研究

工作。

17.1.2 B-R 排序方法应用实例之一:我国华北几个晚更新世动物群的排序

我们知道每个地区动物群的组成是随时间不断变化的,古老属、种的灭绝,新种的出现繁衍。因此动物群的组成反映了它的时代。比较同一地区动物群之间动物种属组成的异同有可能对动物群按照其时代的早晚排序。这里尝试用 B-R 方法对我国华北地区属于晚更新世的 6 个主要的化石动物群进行排序,它们是丁村、许家窑、萨拉乌苏、峙峪、小南海和山顶洞等 6 个动物群。为了对比的方便,将中更新世晚期的古老种和现存的现生种与上述 6 个动物群一起进行排序,这样参加排序的动物群共 8 个。选择了食肉、长鼻、奇蹄和偶蹄 4 目的 36 种动物作分析,其中绝大多数动物能鉴定到种。啮齿目、兔形目和食虫目的动物因各种原因未被选择。因为难以统计每个动物群中每种动物的出现频次,只是统计某种动物是否在某动物群出现。后者属于二元变量。表 17-5a 统计了 36 种动物在 8 个动物群中的分布。

表 17-5a 36 种动物在 8 个华北动物群中的分布,"1"代表存在,"sp"代表只能鉴定到属

	古老种	丁 村	许家窑	萨拉乌苏	峙 峪	小南海	山顶洞	现生种
狼		sp	1	1		1	1	1
狸狼							1	1
狐		sp					1	1
豺							1	1
棕熊		sp					sp	1
洞熊						1	sp	
阿尔泰鼬							1	1
艾鼬							1	1
最后斑鬣狗				1	1	1	1	
猞猁							1	1
豹猫							1	1
香猫							1	1
野猫							1	1
猎豹							1	1
德永象	1	1						
纳玛象	1	1					sp	
诺氏象	1		1	1				
印度象		1					sp	1
梅氏犀	1	1						
披毛犀		1	1	1	1	1		
普氏野马		1	1	1	1			1
野驴		1	1	1	1	1	1	1
野猪		sp	sp	1		1	sp	1
赤鹿		1	1	1	1	sp	1	1
葛氏斑鹿		1	1			sp		

续表

	古老种	丁 村	许家窑	萨拉乌苏	峙 峪	小南海	山顶洞	现生种
北京斑鹿						sp	1	1
肿骨鹿	1	sp						
河套大角鹿		sp	1	1	1			
东北狍子						1	1	1
普氏小羚羊			1	1	1	1	1	1
鹅喉羚		sp	1	1	1			
盘羊				1				1
裴氏转角羊	1	sp	1					
古特赫转角羊	1	sp	1	1				
王氏水牛		sp		1	1	sp		
原始牛		1	1	1	sp		sp	

因为动物种的存在与否属二元变量，不能用17.1节讨论的B-R相似系数来表征实体(动物群)间的相似程度，需要使用二元变量间的匹配系数(见14.3.2)。若某种动物 k 在 i 和 j 两个动物群中都出现或者都未出现，那么定义该动物种在 i 和 j 两个动物群中的匹配系数 $R_{ijk}=1$；如果某种动物 k 仅在 i 和 j 两个动物群中的一个出现，而在另一个动物群没有出现，那么定义该动物在 i 和 j 两个动物群中的匹配系数 $R_{ijk}=0$。由于个别动物只能鉴别到属，相应定义其匹配系数 $R_{ijk}=0.5$ 或0。两个动物群中的总匹配系数是对所有的动物种求和并乘以2，为 $R_{ij}=2\times\sum_{k=1}^{36}R_{ijk}$，其值在0到72之间变化。$R_{ij}$ 是14.3.2节讨论的简单匹配系数。使用简单匹配系数而不使用Jaccard系数，是因为所分析的6个动物群所统计的动物个体数甚多，因此在某个动物群中未观察到某种动物可以认为该动物种的不存在。

计算了8个动物群两两之间的匹配系数后，可以根据B-R排序原则，即利用公式(17-3)和(17-4)对它们进行排序。得到它们的排列次序为古老种→丁村→许家窑→萨拉乌苏→峙峪→小南海→山顶洞→现生种。表17-5b是这8个动物群按照B-R准则排序的相似系数表

表17-5b　按照B-R准则排列后的华北地区8个晚更新世动物群之间的相似系数表

	古老种	丁村	许家窑	萨拉乌苏	峙峪	小南海	山顶洞	现生种
古老种	72							
丁村	44	72						
许家窑	43	50	72					
萨拉乌苏	36	44	61	72				
峙峪	39	45	56	61	72			
小南海	35	38	48	51	54	72		
山顶洞	16	18	24	26	30	41	72	
现生种	12	20	25	28	27	33	60	72

在表17-5中，数值大的相似系数集中在主对角线的邻近，而且从 M_1 到 M_8 的数值是单调下降的，依次为72,53,42.8,36.4,32.3,26,18和12，公式(17-3)的要求得到满足。可

以证明表 17-5 的 $D = \sum_{i=1}^{8} M_i = 292.5$ 是最小值，满足公式(17-4)的要求。这个排序结果能为旧石器考古学家所接受并与已知的测年数据相符(chen et al 1991)。

在完成了动物群的排序后，可以直观地观察每种动物出现或灭绝在时代上与哪个动物群相对应。因为在表 17-5 中动物群是按照 B-R 准则排列的，由表可见，例如最后斑鬣狗最早出现于萨拉乌苏动物群，在山顶洞动物群还存在，但现在已灭绝了。又例如普氏野马从丁村动物群时代一直延续到小南海动物群，但在山顶洞动物群中已见不到它的存在。

从上面的例子可见，Brainerd-Robinson 方法应用于某个地区动物群的排序是可行的，而且能给出与实际测年数据相符合的序列。

17.1.3 B-R 排序方法应用实例之二：江陵雨台山楚墓的排序与分期

裴安平等(1991)曾用黄其煦根据国外程序改编的〈计算机考古年代系列分析系统〉软件(CASA——Computer Archaeological Seriation Analysis)对湖北江陵雨台山 34 座日用器组合齐全和 127 座仿铜礼器组合齐全的楚墓分别进行了年代序列分析。裴安平等所以选择雨台山楚墓作为 CASA 对象，是因为他们认为湖北省江陵博物馆(1984)的原始研究报告《江陵雨台山楚墓》(以下简称《报告》)"资料完整，分期序列明了，对传统考古类型学方法的成功运用获学术界公认"。

以 34 座日用器组合齐全的墓为例，出有 8 种器物，其中的鬲又可分为 4 型。对于 B 型鬲、盂和长颈壶等可进一步分为 I—IV 等 4 式，其他多数器物也能分成 3 式或 2 式。因此可参与比较的器物式别共 26 种。对墓葬而言，有的墓葬间的器物式别的组合是完全一致的，因此 34 座墓葬不同的日用器组合为 22 组。

CASA 接受原《报告》对器物的分型定式，并在此基础上根据器物组成间的相似程度对 22 组组合进行 CASA 排序。裴安平等对 CASA 的排序结果作了分析，注意到出现 2 例倒序，它们是(1)出 III 式 B 型鬲的 M512 处于出 II 式 B 型鬲的几座墓前面，和(2)出 IV 式 B 型鬲的 M483 处于几座出 III 式 B 型鬲的墓前(鉴于篇幅，这里未列出裴文中的"日用器组合齐全墓 CASA 年代分析序列表"，感兴趣的读者可查阅原文)。但倒序的墓葬仅此 2 座，占总墓葬数 5.8%。在对这 2 座墓葬的序列位置作了调整后，B 型鬲的各式不再有倒序现象，而且与原《报告》对 34 座墓葬的分期也一致了。同时他们指出原《报告》中对 22 组日用器组合的排序中，出现倒序的"不止 5 例"。由此他们的结论是"CASA 的年代分析序列""内在逻辑关系较严谨，各种器物型式的排比顺畅，倒排现象被降到最低限度"。本书作者注意到 M512 出 III 式 B 型鬲，又出 II 式长颈壶，但有的墓葬却同时出 II 式 B 型鬲和 III 式长颈壶，因此裴等调整 M512 的位置使 B 型鬲的排列顺畅，必然会导致长颈壶出现倒序。对于 M483 也有同样的情况，这种矛盾现象的存在说明在排序基本合理的情况下出现个别的倒序现象是难免的，也许在排序过程中应进一步考虑 B 型鬲与长颈壶之间哪种器物更具典型性，从而对它们加不等的权重。

裴等对 127 座仿铜礼器组合齐全的楚墓的 32 种组合的排序中仅 M555 出现器物式别的倒序，优于原《报告》所排的序列。M555 原《报告》定为第六期，即最晚期，但它却处于

CASA 序列的中部，两种分期矛盾。该墓出土 II 式 B 型鼎、IV 式敦和 V 式 A 型壶等晚期器物，但也含有 I 式钫。I 式钫仅在三期(早期)的 M472 中有发现。CASA 综合 M555 同时出早期和晚期的器物的现象，将其排在序列的中央。也就是说 CASA 排序方法可能没有充分考虑"以晚期因素确定地层和墓葬时代的类型学"原则。这点可能是目前国内使用多元数量排序方法的共同缺点，它们将所有式别的器物是同等看待的。闫渭清(1991)曾批评朱乃诚(1984)应用概率分析方法于渭南史家墓地的分期中没有适当考虑考古类型学的这个原则。裴等将 M555 在序列中的位置作了调整，并正确指出"CASA 所排序列只具有统计学基础上的逻辑真实性，不能脱离研究者对结果的分析和判断"。此外 CASA 排序在原《报告》将 127 座仿铜礼器组合齐全的墓葬分成四期的基础上，在三四期之间和四五期之间分别插入了两段，而且认为这两段的诸墓葬，"器物组合关系复杂，器物形式多样，具有明显的承前启后性"。

总之，CASA 用于考古实体的排序分期，虽然其基础材料还是基于传统类型学对器物的分型定式，本身还有需改进之处，但它无疑是传统考古学分期方法的一种有价值的补充，特别是当参与排序的实体和描述实体的变量的数量均多，信息量大而又缺乏地层关系时，手工排序的工作量十分庞大，往往会顾此失彼和难免引入"隐含的主观因素"。考古工作者希望寻找一种既能发挥考古学研究方法科学性的长处，又能尽量考虑全面并借助计算机帮助的数学方法应用于考古实体的排序分期。可惜因为某些客观原因 CASA 软件未得到进一步的改进和在其他遗址或墓地中的应用。

17.2　排序与分期的关系——有序实体的最佳分割

对器物、墓葬、遗址乃至考古学文化的分期是考古研究中的重要内容。而 17.1 节所讨论的实体的排序问题，与对实体的分期是紧密相关的。排列有序的一系列实体比将同一批实体粗犷地分为少数几段或几期包含有更多的信息。考古分期可以是建立在排序的基础之上的。如果排列的次序反映时间的早晚，那么在不改变实体的排列次序的前提下，将实体序列划分成若干段就是实现了实体的分期。

有序实体的合理分划或分割需要确定两个问题：(1)分割成几段；(2)怎样确定分割点。因此确定分割的段组数目和寻找分割点也就是有序实体最佳分割所要研究的命题。

17.2.1　有序实体最佳分割的原理和计算过程

首先讨论按单参数排序的实体的最佳分割问题。假设有 n 个实体，它们已根据其某个参数 $x_i(i = 1,2\cdots n)$ 的取值排列有序：

$$x_1, x_2\cdots x_i, \cdots x_n \tag{17-5}$$

x_i 称为排序参数。现要对这个序列做分割。寻找最佳分割点需要先确定将排列有序的实体划分为几段。第一步处理最简单的情况，在将序列划分为 2 段的要求下应怎样寻找最佳分割点的位置。设分割点的位置是 $x^{(2)}$，确定分割点后可以分别计算 2 段各自的离差平方和，并将它们相加得到总离差平方和 $S^{(2)}$。$S^{(2)}$ 的表达式是

$$S^{(2)} = \sum_{i=1}^{x^{(2)}} (x_i - x_1^{(\bar{2})})^2 + \sum_{i=x^{(2)}}^{n} (x_i - x_2^{(\bar{2})})^2 \tag{17-6}$$

式中的$x_1^{(\bar{2})}$和$x_2^{(\bar{2})}$分别为排序参数在第一段和第二段的平均值。显然由于分割点的位置$x^{(2)}$不同,两段各自的离差平方和,以及总离差平方和$S^{(2)}$的值是不同的。确定最佳分割点$x_{\min}^{(2)}$的原则是使得$S^{(2)}$最小,为$S_{\min}^{(2)}$。这个分割原则的含义是使各段内部实体间的差别最小,而不同段的实体间的差别尽量拉大。当然寻找最佳分割点涉及相当的计算工作量,有时可以借助于 SPSS 软件的 K 均值分类程序。

第二步将序列划分为 3 段,相应由两个分割点$x_1^{(3)}$和$x_2^{(3)}$。分别计算 3 段各自的离差平方和后,再相加得到总离差平方和$S^{(3)}$。$S^{(3)}$的表达式是

$$S^{(3)} = \sum_{i=1}^{x_1^{(3)}} (x_i - x_1^{(\bar{3})})^2 + \sum_{i=x_1^{(3)}}^{x_2^{(3)}} (x_i - x_2^{(\bar{3})})^2 + \sum_{i=x_2^{(3)}}^{n} (x_i - x_3^{(\bar{3})})^2 \tag{17-7}$$

式中的$x_1^{(\bar{3})}$、$x_2^{(\bar{3})}$和$x_3^{(\bar{2})}$分别为排序参数在第一段、第二段和第三段的平均值。显然分为 3 段时的总离差平方和$S^{(3)}$也是因 2 个分割点位置的变动而变化的。确定最佳分割点$x_{\min-1}^{(3)}$和$x_{\min-2}^{(3)}$的原则也是使得总离差平方和$S^{(3)}$最小,为$x_{\min}^{(3)}$,$x_{\min}^{(3)}$显然是小于$x_{\min}^{(2)}$的。当然分为 3 段的计算工作量比分为 2 段更大。

可以接着分划 4 段、5 段,要求划分 4、5 段时的总离差平方和$x_{\min}^{(4)}$和$x_{\min}^{(5)}$最小来确定相应的诸最佳分割点的位置。上面的讨论解决了在已知分段数的条件下寻找最佳分割点的问题。但是一个有序排列的实体组最佳应分为几段呢。显然分段的数目愈多,总离差平方和愈小。如果实体系列不分段,总离差平方和最大,而如果将由 n 个实体组成的系列分成 n 段,即每个实体各自组成一段,总离差平方和最小,就等于 0。因此有

$$S_{\min}^{(1)} > S_{\min}^{(2)} > \cdots > S_{\min}^{(n)} = 0 \tag{17-8a}$$

或者

$$\frac{S_{\min}^{(1)}}{S_{\min}^{(1)}} = 1 > \frac{S_{\min}^{(2)}}{S_{\min}^{(1)}} > \cdots > \frac{S_{\min}^{(n)}}{S_{\min}^{(1)}} = 0 \tag{17-8b}$$

对于大多数实际的例子,当分割的段数很少时总离差平方和下降很快,而当分割的段数很多时总离差平方和将基本上趋于 0 而变化很慢的了。可以根据总离差平方和下降的速度来决定应该分划为几段。下面通过实例来显示怎样确定最佳的分割段数。

17.2.2 有序实体最佳分割的实例:河南二里岗期前后陶豆的分期

在第十六章讨论河南地区从二里头期到人民公园期 13 件陶豆的主成分分析时曾指出,这些陶豆依据第一主成分的分布基本上是按时间早晚的排列(见表 16-12 和图 16-8)。我们将表 16-12 的数据抄录于表 17-6,表 17-6 中陶豆是按照它们第一主成分的得分值排序的。

表 17-6 河南地区从二里头期到人民公园期 13 件陶豆的第一主成分值

陶豆编号	考古分期	第一主成分
4	二里头二期	-1.60345
3	二里头四期	-1.39335
2	二里头二期偏晚	-1.26747

续表

陶豆编号	考古分期	第一主成分
1	二里头二期偏晚	-0.80412
b	早于二里岗期上层	-0.1932
5	晚于二里头四期	0.00419
9	二里岗期下层	0.248
d	同 上	0.5017
6	同 上	0.54822
c	同 上	0.5878
a	二里岗期上层	0.61541
8	二里岗期上层	1.20411
7	人民公园期	1.55216

下面对这13件陶豆根据其第一主成分大小序列作最佳分割。

1. 第一步分为2段。按公式(17-6)计算和比较不同分割点时的 $S^{(2)}$ 值,确定分2段时的最佳分割点应在陶豆"1"和"b"间,即二里头4件陶豆为一段,其他各期的陶豆为另一段。两段的中心坐标分别是 -1.267 和 0.563,最佳2段分割的总离差平方和为 $S_{\min}^{(2)} = 2.723$。顺便指出如果不分段,13件陶豆第一主成分得分的总离差平方和为 $S_{\min}^{(1)} = 12.00$。

2. 第二步分为3段。经过计算确定分为3段时的2个最佳分割点为:第一分割点仍在陶豆"1"和"b"间,第二分割点在陶豆"a"与"8"间。即二里头4件陶豆仍处在第一段,6件相当于二里岗下层的陶豆和一件二里岗上层的陶豆在第二段,一件二里岗上层和一个人民公园期的陶豆为第三段。三段的中心坐标分别是 -1.267 和 0.330 和 1.378。分3段时的最小总离差平方和为 $S_{\min}^{(3)} = 1.015$。

如果希望分段更细,例如分为4段,那么将把二里头期的陶豆"1"独立分为一期,其他

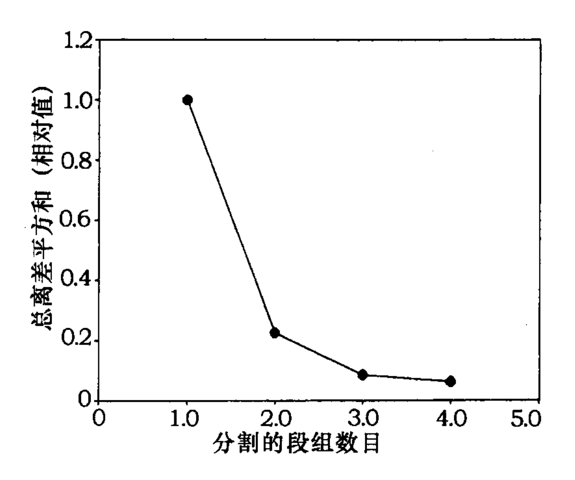


图 17-1　有序实体最佳分割中总离差平方和(相对值)随分割段数的变化图,以河南地区二里岗期前后13件陶豆的分期为例

分段情况不变。分 4 段时的总离差平方和为 $S_{\min}^{(4)} = 0.731$。为了显示总离差平方和随分段数增加而变小的速度可以画一张相应的折线图(见图 17-1),以帮助判断对陶豆序列分为几段较为合适。

由图可见,从不分段到分为 2 段时,总离差平方和的降低极为明显,而将分割段数从 3 段增加到 4 段时,总离差平方和的降低已不是很显著的了。因此这 13 件陶豆分成 2 期或 3 期应该是合适的。上面的例子表明,对按时间有序排列的考古实体进行最佳分割是可以实现对实体的分期的。

17.3 史家墓地的数量方法分期及其相关问题

80 年代中期朱乃诚(1984)和陈铁梅(1985)先后用定量方法对渭南史家墓地的墓葬进行了分期研究,这是在我国最早应用数量方法对考古单位进行分期的尝试。1978 年西安半坡博物馆发表了史家墓地的发掘材料。张忠培于 1981 年利用传统的考古地层学和类型学的方法提出了渭南史家墓地的第一个分期方案。由于朱和陈使用了新的、定量的方法于墓葬的分期,而且他们的分期方案有异于张的方案,由此引起了考古界一场颇有意义的学术争论。参加争论的论文有伊竺(1985),陈雍(1985)和稍晚的刘茂(1989)等,这样对史家墓地共提出了 6 个分期方案,这些方案基于同一批资料,即半坡博物馆 1978 年发表的材料,但各分期方案间是有一定程度的差别的。本节将对下面几方面的问题作讨论:(1)介绍朱和陈进行墓葬分期所使用定量方法的基本思想,取得的结果和优缺点;(2)提出几种判断分期方案异同的定量标准,并在此基础上定量比较这 6 个分期方案间的异同程度;(3)应用地层关系和器物的演化序列检验分期方案,对检验中的某些问题进行探讨。

渭南史家墓地属仰韶墓地。发掘有墓葬 43 座,其中在 37 座中出土 28 种式别的器物共 128 件。器物有钵、罐、葫芦瓶、细颈壶、碗、盂、瓮等 7 种器形,其中后 3 种器物仅在 3 座墓葬中出现了 8 件。表 17-8 列出了钵、罐、瓶和壶等 4 种常见器物 14 种式别共 111 件在 37 座墓葬中的出现分布情况。在这些墓葬间存在 7 组叠压、打破关系,涉及 32 座墓葬。墓葬的 6 个分期方案都是根据墓葬的叠压关系和这些器物在墓葬中的分布作为基础资料来进行的。

表 17-7 史家墓地 37 座墓葬中 14 种常见器物式别的分布表

(墓葬按朱乃诚分期排列)

	钵				罐							瓶		壶	总数	朱分期
墓号	I	II	III	IV	I	II	III	V	VI	VII	VIII	I	II	II		
2										1			1		2	1
10				1		1				1			1		4	1
37				1									1		2	1
9		1						1					1		3	2
12								1					1		2	2
39			1							1			1		3	2

续表

	钵				罐							瓶		壶	总数	朱分期
墓号	I	II	III	IV	I	II	III	V	VI	VII	VIII	I	II	II		
3			1		1					2		1			5	3
4	1					1				1			1		4	3
11			1				1			1		1			4	3
15			1		1					1					3	3
17	1					1				2					4	3
19	1	1						1					1		4	3
21	1									1		1			3	3
23		1								1		1			3	3
28	1					1				2		1		1	6	3
31				1					1				1		3	3
32				1						1		1			3	3
33					1	1				1			1		4	3
36	1					1									2	3
38		1							1				1		3	3
43										1		1			2	3
35											1				1	4
8			1									1			2	5
16	1							1							2	5
18		2							1						3	5
24		1					1				1				3	5
25		1					1				1	1			4	5
1							2			1			1		4	6
5					1				1					2	4	6
14	1								1			1			3	6
22															0	6
26					1							1			2	6
27	1				1				1			1			4	6
29	1														1	6
30								1	1						2	6
34					1				1			1			3	6
42	1						1		1			1			4	6
总和	11	8	5	4	8	5	6	5	9	18	3	14	12	3	111	

17.3.1 概率法分期的基本思想和过程

朱乃诚根据西安半坡博物馆(1978)发表的基础资料,用他称之为概率法的方法,提出了一个史家墓地的分期方案,这是在我国第一次应用定量方法于墓葬的分期。该方法的基本思想是朱自己发展的,其决定墓葬期别的步骤如下。

1. 第一步是根据墓葬叠压关系和器物的共存关系定出早晚两期的典型墓葬和典型器物。在7组有叠压关系的墓葬组中,排除了虽有叠压关系但又出有相同式别器物的墓

葬后,初步选出 M15,M10,M37 和 M39 为早期墓葬,M5 和 M14 为晚期墓葬。但又因为 M15 和 M5 中出有 3 种相同式别的器物,它们不能被认为是典型墓葬。在 M10,M37 和 M39 中共同出现的器物式别很多,又以 M10 出的器物式别最多,这样定 M10 为典型的早期墓葬,其中所出的钵 IV,罐 II,VII 和瓶 II 被定为 4 种典型早期器物。晚期墓葬的代表是 M14,其中所出的钵 I,罐 VI 和瓶 I 则定为晚期典型器物。确定早、晚期典型器物是概率法进行墓葬分期的关键步骤。

2. 第二步是器物的分期。依据每种式别的器物(以下简称每种器物)在墓葬中与典型器物之间的共存关系的频繁程度,定出每种器物与早、晚期典型器物共存的概率,称为早期或晚期组合概率。朱文认为早期的器物会与钵 IV 等 4 种典型早期器物共存的机会多,而与钵 I 等 3 种典型晚期器物共存的机会少,相应其早期组合概率高而晚期组合概率低。对于晚期器物,则情况会反过来,他们的晚期组合概率将高于早期组合概率。因此朱使用每种器物的早期组合概率和晚期组合概率间的比值来确定该种器物的期别。下面以罐 I 为例,说明计算组合概率和进行分期的过程。由表 17-8 已知钵 IV,瓶 II 和罐 II,VII 等 4 种早期典型器物在 37 座墓葬中共出现了 34 次(如某种式别的器物在同一墓葬中出现 2 件,仍作为出现一次计入),而罐 I 与这 4 种早期器物共在 6 个墓葬中共存,因此罐 I 与早期典型器物的组合概率为 $P_e = \frac{6}{34} = 0.177$。同样方法可以计算罐 I 与晚期典型器物的组合概率为 $P_l = \frac{12}{35} = 0.343$,式中的"12"系罐 I 与 3 种典型晚期器物共存的次数,而"35"为 3 种典型晚期器物总共出现的次数。每种器物根据它的早、晚期组合概率的比值被定为早期器物,中期器物和晚期器物。对于罐 I,其$\frac{P_e}{P_l} = \frac{0.177}{0.343} \approx 0.5$,晚期组合概率约为早期组合概率的 2 倍,它应定为晚期器物。有的器物,例如对于钵 II,它的 2 个组合概率的比值为 1.4,差别不大,应定为中期器物。在朱的分期工作中,对于上述的 7 种典型器物也同样要计算它们的组合概率,重新分期。例如对于钵 IV,可计算得到 $P_e = 0.294$, $P_1 = 0.057$,其早期组合概率约为晚期组合概率的 5 倍,当然应定为早期器物,从而验证将钵 IV 定为典型早期器物是合适的。这样每种器物的组合概比值决定了该器物应定为早中晚 3 期中的哪一期。朱分别称它们为第 1,2 或 3 组器物。

3. 第三步是根据每座墓葬中出现的器物的分期情况对墓葬分期。朱作规定如下:如果

墓葬中只有第 1 组器物,	墓葬定为 I 期
墓葬中同时有第 1,2 组器物,	墓葬定为 II 期
墓葬中同时有第 1,2,3 组器物,	墓葬定为 III 期
墓葬中仅有第 2 组器物,	墓葬定为 IV 期
墓葬中同时有第 2,3 组器物,	墓葬定为 V 期
墓葬中仅有第 3 组器物,	墓葬定为 VI 期

这样完成了对 37 座墓葬的分期。这个分期方案可以在表 17-7 中看到。应该认为,朱乃诚的分期方案基本上是成功的。因为这个方案能通过地层关系和器物发展序列的检验。在已知的 26 对墓葬叠压打破关系中,在朱的分期方案中仅有 M3 - M34 和 M5 -

M30 两对墓葬分期关系与实际叠压关系矛盾，地层关系的符合率还是比较高的。各种器物按墓葬分期的排序基本上体现先后有序的规律，先出现的器物，一般也较早的消失；出现较晚的器物，能延续到最后。在表 17-7 中能观察到各式别器物演化的规律。

本书作者(1985)在充分肯定朱文分期研究的同时，指出朱文不应该将瓶 IV、盂、碗和瓮等仅在个别墓葬出现 1 次的孤种器物纳入考虑之中，也不应该对像 M22 和 M30 等仅出现一件器物的墓葬进行分期。这类孤种器物和仅含单器物的墓葬随机性很大，不仅对它们的分期不可靠，而且还有可能干扰其他墓葬的正确分期。例如第 4 期仅含有一座墓葬(M35)，而该墓出罐 VIII、盂和碗。其中罐 VIII 仅出现 2 次。而盂和碗为孤种器物。如果将 M35 排除而不予考虑，那么分期方案中的第 IV 期被撤消，从而器物按墓葬分期的演化规律更为清晰。

对于“概率法”最严重的批评来自闫渭清(1991)，闫正确地指出概率分期方法没有充分地考虑考古地层学的一个基本原则，即在晚期的考古单位中可以出现早期遗物，考古单位的时代应该由其中最晚的遗物来确定。所幸的是，朱在确定早、晚期典型墓葬和典型器物时是默认了这个原则的，他将有相互叠压关系，而同时又有共存器物的墓葬排除在典型墓葬之外。正如前面所言，由地层关系所定的早期墓葬 M15 与晚期墓葬 M5 间有相同的器物，因而它们未被选择为早、晚期典型墓葬。因此朱的这个“忽略”并没有影响他分期工作的前面二步，即没有影响对早、晚期典型器物的正确选择和器物的分期。问题出在分期工作的第三步，即(3)中提出的按墓中出现 1,2 和 3 组器物组合的情况来定墓葬期别的规则是有悖于“由考古单位中最晚的遗物来确定单位的时代”这个基本原则的。但应该说，因没有充分考虑这个基本原则所导致分期方案的错误应该是局部的，也许需要对某些被定为早中期的墓葬重新考虑它们的分期。

17.3.2　聚类方法分期的思想和过程

本书作者(陈铁梅，1985)曾尝试使用聚类方法于史家墓地的墓葬分期，其分期过程分为 7 步。

1. 第一步对原始资料进行了筛选。将仅出现 2 次或单次的偶见器物和仅含有 1 种器物的“贫瘠”墓葬排除。因此仅对 32 座墓葬进行分期，这些墓葬中包含有 4 种器类的 13 种式别，它们是钵Ⅰ、Ⅱ、Ⅲ、Ⅳ式，罐Ⅰ、Ⅱ、Ⅲ、Ⅳ、Ⅴ、Ⅶ 式，瓶Ⅰ、Ⅱ式和Ⅱ式壶。

2. 对墓葬的器物组成作定量描述。每座墓葬的属性是根据它的器物而确定的，13 种器物在该墓中出现哪几种。这样每个实体(墓葬)的属性由 13 个按一定次序排列的二元变量的取值所决定。若某种器物出现，则相应变量取值为 1，若未见某种器物，则相应变量取值为 0。13 种器物是排好固定的顺序的(器物排列的顺序见表 17-8 的第一行，但不包括罐 VIII)，每个墓葬的属性就可以用一组 13 个二元变量来表示了。例 M2 为(0,0,0,0,0,0,0,0,0,1,0,1,0)，因为出有罐 VII 和瓶 II；而 M10 为(0,0,0,1,0,1,0,0,0,1,0,1,0)，因墓中出有钵 IV，罐 II，罐 VII 和瓶 II。32 座墓最终被一个 32 行、13 列的原始数据的矩阵所描述。每行代表一个墓葬，反映它出有哪几种器物；而每列代表一种器物，反映它在哪几座墓葬中出现。

3. 墓葬间器物组成相似程度的定量描述。聚类分析首先需要定义和计算实体之间

的相似系数或相异系数。陈的分期工作应用 Jaccard 系数描述墓葬间异同的程度。在第十四章 14.3.2 节中曾给出 Jaccard 系数的定义和计算方法。Jaccard 系数的定义为 $S = a/(a + b)$。a 为二个墓中都出现的器物的种类数，称为(1,1)匹配。b 是仅在二墓中的一座墓中出现的器物种类数，称为(1,0)或者(0,1)匹配。在计算 Jaccard 系数时，在二个墓中都不出现的器物种类是不起作用的。比如说在上述的 M2 和 M10 之间，共同出现的器物种类数是 2(罐 VII 和瓶 II)，仅在一个墓出现的器物种类数也是 2(钵 IV 和罐 II)，那么这二个墓之间的 Jaccard 系数应该是 $2/(2+2) = 0.5$。显然这样定义的 Jaccard 系数总是在 0 与 1 间变动。二个墓的器物组成越接近，它们间的 Jaccard 系数就越接近 1，反之，如果二个墓的器物组成差异越大，它们间的 Jaccard 系数就越接近 0。所以 Jaccard 系数是表征墓葬间器物组成相似程度的一种度量。32 座墓葬每两两间都计算 Jaccard 系数，得出一个 32 行 32 列的 Jaccard 系数的矩阵。这是一个对称的方阵。

4. 根据 Jaccard 系数矩阵对墓葬进行聚类。建立了 Jaccard 系数矩阵后，用 1 减去所有的 Jaccard 系数，可得到墓葬间的相异系数矩阵，并根据相异系数矩阵采用均值聚类法对墓葬进行聚类。14.4 节曾对该方法的原理作了详细讨论，聚类程序大致如下。在墓葬间的相异系数矩阵中，选取最小的一个系数值，这个系数所处的行和列对应的 2 个墓葬，它们应是相异程度最小的一对墓葬，其器物组成应该最接近。将它们聚为一组。然后以其他各墓与这二个墓的相异系数的平均值作为它们与新的合并组之间的相异系数，从而得到一个新的相异系数矩阵。这个新的矩阵是 31 行 31 列，即行数和列数均比原始的相异系数矩阵少 1。这样一步步地把 32 座墓葬按随葬器物组成间的相近程度逐组归并成类，最后得到一个表示各墓葬器物组成间相互接近程度的“聚类树枝状图”。

5. 墓葬的分组与分期。根据不同的聚合水平，即每一步聚类时的相异系数值，可以在树枝状图上将实体分成 2 组、3 组或更多的组。每组组内各实体间的性状应比较接近，而不同组的实体间的性状应该相差较大。陈的分期方案中把 32 座墓分成了 4 组。需要指出，上面已完成的仅是对墓葬的分类，还没有达到分期的目的。因此需要根据墓葬间的已知的叠压关系将 4 组墓葬放在时间标尺上，即将 4 组墓葬按年代的早晚排列。从而实现了墓葬的初步分期。

6. 用层位关系检验初步分期方案和对个别墓葬期别的调整。陈的分期方案包括 32 座墓葬，其中有 25 座墓葬相互间存在叠压或打破关系。每个分期方案都必须与每一组墓葬间的叠压打破关系相符，不能违背。对上面的初步分期方案作检验，结果是其中有 17 起是晚期打破或叠压早期的，有 4 起是同期墓葬间存在打破关系的，但也有 4 起是晚期墓葬被早期墓葬打破的。总体上说，陈的初步分期方案与地层关系相吻合的比例还是比较高的。但是晚期墓葬被早期墓葬打破的情况是不能容忍的，需要对这 4 组其地层叠压关系矛盾的墓葬的期别作调整，以得到最后的分期方案。

7. 器物演化序列的检验。考察四期墓葬中所表现出的器物演化模式，也是检验分期方案是否合理的一种标准。对所选的 13 种器物中，有 8 种器物都表现出“产生、发展和消失”的演化规律，但有 4 种器物(钵 I、钵 II、罐 I 和罐 VI)在四期中都有出现，还有一种器物是例外(罐 III)，在二期和四期墓葬中出现，三期却未见，罐 III 一共只在三个墓葬中出现，属于不常见的器物，也许可以作为例外或偶然现象来处理。因此总体来说，陈文“对史家

墓地的分期经受了层位和器物的检验"(腾铭予(2001),计算机考古讲义,吉林大学内部应用讲义)。关于用器物演化序列来检验史家墓地墓葬分期的问题在下一节将作进一步讨论。

8. 刘茂(1989)对陈文的批评有下列两点。(1)刘茂重复了陈文所执行的聚类过程,注意到在聚类的过程中曾出现了2个相等的最大 Jaccard 系数,选择哪个系数进行下一步聚类会影响后面的聚类过程,也就是说,聚类的结果可能不是唯一的。刘的意见是值得注意的,但对于史家墓地的这个实例,不同的选择虽会影响某些墓葬的聚类次序,但作为聚类分析最终结果的树枝状图差别并不大,因为聚类树枝状图反映的是墓葬间器物组合的总体异同情况。(2) 刘茂还批评陈和朱的定量方法中混淆了器种和式别的概念,把同种器物的不同式别作为不同种的实体对待。这点意见也是正确的。但是将器种和器物的式别作为两个层次的实体进行处理,对于目前的定量分析方法是有困难的,各种定量方法都将不同器种的各个式别作为同一层次的"类"来处理。目前所能做的是,对分期方案中"类"的演化模式与器物式别演化的逻辑序列和地层序列进行比较,作为检验定量方法分期方案的标准之一。

17.3.3　比较史家墓地六个分期方案间异同程度的数值度量

前面提到,对史家墓地至今已提出有6种分期方案。最早是张忠培用传统考古方法的分期,朱和陈各自用定量方法提出2个分期方案,此外伊竺(1985)和陈雍(1986)用传统的方法,刘茂用传统方法并参考了概率法的模式也各自提出了自己的分期方案。这6种分期方案的材料依据是完全一样的,都是西安半坡博物馆1978年发表的简报的材料,即37座墓中出土的26种器物共123件和7组墓葬间的叠压、打破关系。但这6种分期方案却并不是完全一致,有的方案之间还差别较大,个别墓葬在不同的方案中可以分别分到早期或晚期。在6位分期方案的提出者之间曾相互争论和批评,但并不能客观地判断出孰是孰非,哪个方案更符合实际。每一个分期方案本身基本上是内洽的,也不悖于已知的7组墓葬间的叠压、打破关系。因为有11座独立的墓葬,它们与7组叠压关系无关,而且这7组叠压系统相互之间的地层关系也不明确,因此虽然每个分期方案必须符合已知的7组叠压关系,但反过来符合7组叠压关系并不能保证分期方案的正确,即与地层关系相符是分期方案合理的必要条件,但不能作为充分条件。为了判断哪个分期方案更合理,需要分析比较各分期方案给出的器物演化序列,考察哪个分期方案给出的器物演化序列更符合实际。但在分析比较器物的演化序列前,先探讨怎样判断两个分期方案之间相似或相异的程度,需要寻求一个客观的定量标准来度量。

本书的作者认为朱和陈的分期方案接近,而与张的分期方案有差别,但陈雍(以下简称雍)却认为朱和陈虽都用的是数学方法,但"由于两文采用的具体方法不同,分期的结果也有较大出入"。"接近"和"较大出入"都是模糊的度量,看来比较两个分期方案之间的相似或相异程度需要一个客观的,定量的标准。可以定义好几种衡量两个分期方案之间的相似或相异程度的定量标准。第十一章的11.2节和11.3节曾分别计算了陈和张的两个分期方案之间的 Gamma 等级相关系数和 Kendall's tau-b 等级相关系数,用这两个等级相关系数来定量表述这两个分期方案异同的程度。除等级相关系数外,还可以利用公

共信息系数或者结合系数(coherence coefficient)等来定量表述分期方案的异同程度。下面我们介绍一种也许不是十分严格,但却简单明了、并容易为考古工作者了解和接受的方法,来定量表述两个分期方案的异同程度。

考虑两种方案共同对 n 个墓葬进行了分期,为简化说明过程,假设两种方案都把 n 个墓葬分成早中晚 3 期。首先定义每个墓葬的“分期变量”如下,规定:它对于早期墓葬取值为 0,对于中期墓葬取值为 5,对于晚期墓葬取值为 10。第二步计算每个墓葬在两个分期方案比较中的“得分”,定义为该墓葬的两个分期变量的差值的绝对值,即如果某墓葬在两种方案中的分期是一致的,得分为 0,如果差一期得分为 5,差二期(即一个方案定为早期,另一方案定为晚期)得分为 10。这样定义计算的每个墓葬的“得分”值反映了两个分期方案对该墓葬所定的分期位置的相离程度。把全部 n 个墓葬的得分加起来求和,再将得分和除以墓葬数 n,就得到两个分期方案中 n 个墓葬的平均得分。显然,如果两种方案对每个墓葬的分期都是一致的,那么平均得分应该为 0,反之如果每个墓葬在两种方案中的分期都是完全矛盾的(这是一种极端的情况,只是理论上的可能性,每个墓葬在一个分期方案中如定为早(晚)期,则在另一个方案中定为晚(早)期,没有被定为中期的墓葬),那么平均得分为 10。墓葬的平均得分总是在 10—0 之间波动,其大小应该可以作为两个分期方案之间相异或相似程度的定量标准,数值越小表示两个方案越接近。我们称平均得分为两个分期方案之间的相异系数。

上面是假设两个分期方案均将墓葬分成早中晚三期的情况。如果两个方案分期的期段数不一致或不是分为 3 期,依然可以用同样的原则计算墓葬的“分期变量”和“得分”,计算得到方案间的相异系数。例如某个分期方案将墓葬分为 4 期,那么早晚期墓葬的“分期变量”仍分别定义为 0 和 10,但对第二和第三期墓葬的分期变量定义为 3.33 和 6.66。后面可以按照同样的原则计算每个墓葬对于两个分期方案的“得分”和全部墓葬的平均得分,即两个分期方案之间的相异系数。

按上述原则计算得到,对于朱陈两分期方案的相异系数 $S = 1.94$,对于朱张两方案 $S = 2.66$。6 个分期方案两两间计算得到 21 个相异系数值,其中包括每个方案自己对自己的相异系数,后者应该是等于 0 的。将这些相异系数值列表显示如下。

表 17-8 史家墓地六个分期方案间的相异系数表

	朱	陈	伊	张	雍	刘
朱	0	1.94	2.81	2.66	2.45	2.56
陈		0	2.50	3.33	3.22	3.44
伊			0	2.59	3.14	3.68
张				0	1.21	1.92
雍					0	2.53
刘						0

由表 17-8 可见:(1)6 个分期方案间总体上存在相当多的共性,即使最大的相异系数值为 3.68,也处于 0-10 间隔接近于 0 的一端;(2)在诸分期方案中,朱陈间,张雍间和雍刘间相对比较接近,相应的 3 个相异系数均小于 2。但为了清楚地显示 6 个方案间异同的总体情况,根据表 17-9 所列出的相异系数矩阵作均值聚类分析。均值聚类分析的结果由

聚类树枝状图所示(图 17-2)。由图 17-2 可见,张和雍两个分期方案最相接近,它们在相异系数 1.21 的水平上首先聚为一组。朱与陈也较接近,在 1.94 的水平上聚类。第三步是刘和张雍混合组聚类,聚类水平是 2.2。接着第四步是伊和朱陈混合组聚类,聚类水平是 2.65。最后 2 个混合组在相异系数 3.2 的水平上聚为一大组,包含了全部 6 个分期方案。这张聚类图的聚类次序和聚类水平形象而且定量、客观地描述了 6 个分期方案相互间的异同程度。聚类分析的结论是:6 个分期方案大致可分成两组,一组包括张,雍和刘,另一组包括朱,陈和伊,下面我们称之为张群和朱群。在张群中以张和雍的两个方案最为接近,在朱群中以朱与陈的两个方案最为接近。

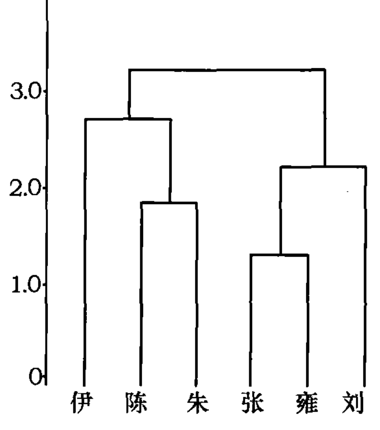


图 17-2　史家墓地六个分期方案间的聚类图

顺便指出,使用 GAMMA 等级相关系数等其他的描述分期方案间的相似系数,可以得到与图 17-2 大致相同的树枝状聚类图,都是分成张雍刘和朱陈伊两组,只是在聚类水平上稍有差异。因此可以说,图 17-2 所示的聚类图客观地反映了 6 个分期方案间的异同关系。本小节的讨论也使我们看到定量地评估两个分期方案之间的相似或相异程度,比通常直观、模糊地判断有明显的优点,而且实现起来也并不难,可以用简单的数学方法来实现。

17.3.4　根据器物的演化序列对史家墓地几个分期方案的检验

前面聚类分析表明,史家墓地的 6 个分期方案基本上可以分成张群和朱群两组。由于墓葬间叠压、打破关系的墓葬组较少,难以判断哪组方案更接近实际。分析器物的演化序列也是检验分期方案合理程度的一个指标。本小节将从器物的演化序列的角度来比较分析朱群和张群的分期,并分别以朱乃诚和陈雍的两个方案作为朱群和张群的代表。朱乃诚方案中 14 种常见器物式别的演化序列已在表 17-7 中列出,这里同样列出这 14 个式别的器物在陈雍方案中的演化序列(见表 17-9)。该表中的墓葬是按陈雍的分期方案排列的,但表中同时标出每座墓葬在 6 个分期方案中的期别,便于相互比较。

表 17-9　史家墓地 37 座墓葬中 14 种常见器物式别的分布以及每座墓葬在 6 个分期方案中分期表(墓葬按陈雍的分期方案排列)

墓号	钵				罐							瓶		壶	分期					
	I	II	III	IV	I	II	III	V	VI	VII	VIII	I	II	II	雍	朱	陈	尹	张	刘
10				1		1				1			1		1	1	1	1	1	1
37				1									1		1	1	1	1	1	1
31				1					1				1		1	3	1	1	1	2
2										1			1		1.5	1	1	1		1

续表

墓号	钵				罐							瓶		壶	分期					
	I	II	III	IV	I	II	III	V	VI	VII	VIII	I	II	II	雍	朱	陈	尹	张	刘
39			1							1			1		2	2	1	1	1	1
3			1		1					2		1			2	3	3	2	1	1
11			1				1			1		1			2	3	2	2	1	1
15			1		1					1					2	3	2	1	1	1
32				1						1		1			2	3	2	1	1	1
8			1									1			2	5	3	2	1	1
34					1				1			1			2	6	4	2		3
9		1						1					1		3	2	2	1	2	3
23		1								1		1			3	3	3	2	2	2
38		1							1				1		3	3	1	1	2	2
18		2							1						3	5	2	2	2	3
24		1					1				1				3	5	2	1	2	2
25		1					1				1	1			3	5	4	2	2	3
12								1					1		3.5	2	2	1		3
35											1				3.5	4		1	2	
1							2			1			1		3.5	6		2	3	2
4	1					1				1			1		4	3	1	1	3	2
17	1					1				2					4	3	2	2	3	
19	1	1						1					1		4	3	2	1		3
21	1									1		1			4	3	3	2	3	2
28	1				1					2		1		1	4	3	3	2	3	2
36	1					1									4	3	1	2	3	2
16	1							1							4	5	2	2	3	3
5					1				1					2	4	6	3	2	3	3
14	1								1			1			4	6	4	2	3	3
22															4	6		2	3	3
27	1				1				1			1			4	6	4	2	3	2
29	1														4	6		2	3	3
30								1	1						4	6		2		3
42	1						1		1			1			4	6	4	2	3	
33					1	1				1			1		?	3	1	1		1
43										1		1			?	3	3	2		1
26					1							1			?	6	4	2		2
和	11	8	5	4	8	5	6	5	9	18	3	14	12	3	34	37	32	37	29	34

史家墓地的主要器物类是葫芦瓶、钵和罐。根据表 17-7 和表 17-9，对这 3 种器形以及壶等 14 种式别在朱、雍两方案中的分期位置汇总于下面表 17-10。

表 17-10 朱乃诚和陈雍对 14 种式别的器物分期情况的对比表

器物种类和式别	数量	朱乃诚分期方案(共分 6 期,但第 4 期仅 M35 一座墓葬)	陈雍分期方案(共分 4 期,但少数墓葬被置于 1,2 期间或 3,4 期间)	朱雍比较
钵 I	11	3—6 期	全为 4 期	均晚
II	9	2—5 期,以 5 期为多	分在 2—4 期	均偏晚
III	5	2—5 期	全在 2 期	朱中雍偏早
IV	4	1 期和 3 期	1—2 期	均早
罐 I	8	3—6 期	2—4 期	均偏晚
II	5	1—3 期	各期均见	朱早
III	6	3—6 期	2—4 期	均偏晚
V	5	2—6 期	3—4 期	朱中晚,雍晚
VI	9	3—6 期	除 1 期 M31,2 期 M34 外,主要在 4 期	均偏晚
VII	18	除 6 期的 M1 外,全在 1—3 期	各期均见	朱早
VIII	3	5 期	主要在 3 期	均中偏晚
瓶 I	14	3—6 期	2—4 期	均偏晚
II	11	除 6 期的 M1 外,全在 1—3 期	各期均见	朱早
壶 II	3	3—6 期	4 期	均晚

考察表 17-10,朱和雍两个分期方案对钵罐瓶 3 种器形各式的分期未见严重的分歧,各式别的分期在两个方案中基本是相符的,未见任何一种器物式别在一个方案中定为早(晚)期而在另一方案中被定为晚(早)期。但是更细致的观察能注意到:(1)雍方案对 4 种式别的钵,从 IV→III→II→I 清楚地对应由早到晚的演化序列。在朱的分期方案中,虽然各式钵也显示了同样的演化规律,但每种式别所跨越的期段较长。如 III 式钵在朱方案中跨越 2-5 期,而在雍方案中均集中于 2 期。(2)朱将 II 式瓶和 I 式瓶清楚地分为早晚两期,而雍虽也将 I 式瓶定为中晚期,但 II 式瓶却从早到晚均有出现。(3)类似 II 式瓶的情况还有罐 VII 和罐 II,朱把它们都定为早期的器物,但在雍的方案中这 2 种式别的罐与 II 式瓶的分布情况相似,都是从早到晚均有出现,被认为是没有分期意义的器物。我们注意到在朱的分期工作中是将这 3 种器物定为典型早期器物,它们在朱的分期工作中起到"基础性"的作用。在陈的分期方案中这 3 种式别也被定为早期或偏早。判断这 3 种式别(罐 II、罐 VII 和瓶 II)是否确实为早期器物,有助于分析朱雍这两个分期方案哪个更为合理。但史家墓地是"单一的典型的文化内涵的遗址",其延续的时间可能不会太长,每种器物类别其式别变化的规律不是十分明显,个别器物式别有可能在史家的早晚期都沿用。目前文献中也未见有关于对史家墓地器物演化的逻辑序列的分析。因此在史家墓地这个孤立的"小系统"中,难以分辨哪种器物演化模式,哪个分期方案更符合实际,也许应把史家墓地放在一个更大的时空环境中加以观察认识。史家墓地的原简报认为史家"介于半坡类型与庙底沟类型之间的一种文化遗存"。因此也许应跳出史家墓地的"小系统",把史家的钵、罐、和瓶等器物放在该地区其他半坡类型和庙底沟类型遗址出土的器物这个"大系统"中来考察比较,能更清楚地认识史家器物的演化规律。张忠培曾进行了这方面的比较,并提出史家的 I 式钵,V 式和 VII 式罐等"与元君庙三期的器物雷同"。

但在朱和雍对史家的分期中都把I式钵和V式罐定为晚或偏晚，对于VII式罐朱定为早期，雍认为各期均出现。因此与元君庙三期的器物比较只能作出史家“当早可至元君庙墓地三期”的推论，而不能帮助判断朱雍两个分期哪个更符合实际。另一种常见器物是瓶，在史家的I式和II式瓶间，相比后者与半坡的葫芦瓶在形态上更相似，因此朱将II式瓶定为早期也许是合适的。总之把史家墓地放在一个更大的时空环境中加以观察的做法应该有助于对目前分期方案的改进。

17.3.5 关于墓葬分期中的几个问题

1. 怎样对待出现次数很少的“偶见”器物和出土器物很少的墓葬。陈铁梅用聚类方法于史家墓地分期时将出现次数不超过2的偶见器物和仅含有1种器物的“贫瘠”墓葬排除在外，在37座墓葬出土的9种器类28种式别中，仅考虑4种器类的13种式别的分布对32座墓葬进行分期。排除“偶见”器物和“贫瘠”墓葬是聚类分析中为正确计算墓葬间的相似系数所要求。但是在所谓偶见器物中有可能存在与墓地的前后文化类型有联系的“典型器物”。如果排除典型器物显然是不适当的。因此在建立初步的分期方案的基础上，应该重新审查被排除的器物和墓葬，考察是否应纳入作为分期的参考标准。

当然反过来排除“常见”器物，而只考虑偶见器物对分期的作用也是不妥当的。“常见”器物出现次数多，有更大的可能在初步分期的各期墓葬中都可见到，因此被当作无分期意义的器物而被排除。例如在张忠培的分期中仅考虑III、IV、V、VI和VII等5种式别的瓶，但这5种瓶在史家共出现6次，而共出现25次的I和II式瓶却没有得到充分地重视。同样7种式别的罐共出现53次，在张的分期中仅考虑了V和VIII式罐，这2种罐仅出现了8次。这也许说明为什么在张的方案中，各式罐和瓶的演化模式不如钵的演化模式那么清楚。

2. 器类和式别的关系。刘茂正确地批评朱、陈等数量分类方法把各类各式器物统视为彼此独立的“种”来处理，特别指出只有器类才可以分为“常见”和“偶见”，某种器类的某种式别出现次数少，不能认为该式别所属器类为“偶见”器类，而且某“偶见”式别之所以为偶见，可能是因为它处于该器类演化的某个特殊环节，因此器物的该“偶见”式别仍有分期意义。器物的类别和式别的层次是不同的，式别是同一器类演化的不同阶段。可惜目前的数量方法只能对同层次的实体进行分类和排序，对实体的诸属性作为同层次的属性处理。因此数量方法的分期结果，除要经受地层关系的检验外，还应检验每种器物不同式别的分期能否符合器物形制演化的逻辑序列，经受该类器物在更大的时空环境中演化情况的考察。

3. 前面已提到，闫渭清(1991)批评朱的分期方法中对“晚期墓葬中出现早期器物的可能性”这个考古地层学的原则未作充分考虑(在确定早、晚期标准器物时朱是考虑了这个原则的)。实际上朱在确定某些墓葬的期别时也注意了这个原则，例如将出II式瓶和VII式罐等早期器物的M1仍安排在晚期，因为其中出有III式罐和VI瓶等晚期器物。应该指出陈的分期方案在计算墓葬间的相似系数时，也没有充分考虑这种可能性。现在看来陈的分期工作显得不够细，对聚类的结果，仅根据地层关系调整了4座墓葬的期别就作为分期方案提出，没有从器物演化的角度作检验和由此对某些墓葬的期别作进一步的

调整。在陈的方案所分析的 13 种器物式别中,有 4 种式别(钵 I、钵 II、罐 I 和罐 VI)在 4 期中均出现,类似于张方案中瓶 I、瓶 II、罐 I、罐 II 和罐 VII 的分期情况。这种多种式别在各期均出现的现象,仰或是符合实际情况(史家墓地的延续时间不可能太长),还是因为分期方案不细致所至。值得深入考虑。

4. 分析 6 个分期方案间,主要是朱陈和张雍二组间的共同点和相异点对进一步更符合实际地安排史家墓地墓葬和器物式别的期别应该是有意义的。首先考察共同点,分析表明 6 个方案均将 M2,M10,M11,M37 和 M39 等墓葬定为早期或偏早,这主要是因为各方案均同意将 IV 式和 III 式钵定为早期或偏早期的器物。朱陈将 II 式和 VII 式罐以及 II 式瓶定为早期或偏早的器物,而张雍方案认为这 3 种式别的器物在史家的各期均可能出现,这也不会引起对墓葬分期的明显矛盾。6 个方案均将 M14,M22,M27,M29,M42 定为晚期或偏晚的墓葬,相应的多数方案均认为 I 式、II 式钵,I 式、III 式和 VI 式罐以及 I 式瓶为晚期或偏晚的式别,有的方案认为其中某些式别从早到晚均有出现。再考察相异点,不同的方案的相异点主要反映在墓葬的分期上,有的墓葬在一个方案中定为早期,而在另一个方案中却定为晚期,例如 M4,M8 和 M38 等。在对器物式别的分期方面,各方案间却未见似墓葬分期那样的完全相悖的情况(见表 17-11)。现分析 M4 的情况,该墓陈定为早期,朱定偏早而张雍定为晚期,明显有矛盾。该墓出土有 II 式和 VII 式罐,II 式瓶以及 I 式钵。朱陈认为前 3 种式别是早或偏早的器物,I 式钵是偏晚的器物,朱陈未考虑"应以墓葬中最晚的器物来定墓葬的年代"这个原则,仅考虑 M4 中偏早的器物占大多数,从而定 M4 的年代也应偏早。从这个问题分析,朱陈所采用的定量分析方法有需改进之处。M38 的分期矛盾也缘于此,地层上 M4 叠压 M38,但这 2 座墓与其他墓葬间没有叠压或打破关系。M8 的分期情况是:张雍定为早期或偏早,朱陈定为偏晚。该墓出土 III 式钵和 I 式瓶。朱陈和张雍均认为前者为偏早的式别而后者为偏晚的式别。从地层关系看 M8 叠压 M40,后者未出任何器物。由此分析,张雍将 M8 定为早期似有点勉强。

前面的分析讨论,显示了数量方法应用于考古单位的分期的可能性,也揭示了数量方法和传统方法应用于史家墓地分期中共同点和相异点,各自的成功和不足之处。显然这两类方法均不够完善,但是可以相互补充,而不是互相排斥的。作为传统考古分期方法的补充,数量方法也需要改进和发展。数量方法与传统方法的互补性应当可以推广到考古学研究的很多方面。这个观点是本书的重要结论之一。

参考文献

蔡莲珍,仇士华,1999,《贝叶斯统计应用于碳十四系列样品年代的树论校正》,《考古》1999年3期。

陈建立,2000,《数学分析方法在考古学中的应用》,《中原文物》2000年1期。

陈靓,2002,《新疆尉犁县营盘墓地古人骨研究》,《边疆考古研究》第1辑,科学出版社。

陈山,2002,《喇嘛洞墓地颅骨种族类型研究》,《边疆考古研究》第1辑,科学出版社。

陈铁梅,1983,《用Brainerd-Robinson方法比较华北地区晚更新世几个主要动物群的年代顺序》,《人类学学报》2卷2期。

陈铁梅,1985,《多元统计方法应用于考古学中相对年代研究——兼论渭南史家墓地三种相对年代分期方案的比较》,《史前研究》1985年第3期.

陈铁梅、何驽,1989,《计算机技术对二里头二期至人民公园期陶豆的分期》,《考古学文化论集》(2),文物出版社。

陈铁梅,1990,《中国新石器墓葬成年人骨的性比异常问题》,《考古学报》1990年4期。

陈铁梅,1991,《我国古代居民颅骨的聚类分析和主成分分析》,《江汉考古》1991年4期。

陈铁梅,1993,《考古学中的定量研究》,《考古与文物》1993年6期。

陈铁梅,Rapp G.,荆志淳,何驽,1997,《中子活化分析对商时期原始瓷产地的研究》,《考古》1997年7期。

陈铁梅,Rapp G., 荆志淳,2003,《商周时期原始瓷的中子活化分析及相关问题讨论》,《考古》2003年7期。

陈雍,1986,《史家墓地再检讨》,《史前研究》1986年3-4期。

赤峰中美联合考古研究项目,2003,《内蒙古东部(赤峰)区域考古调查阶段性报告》,科学出版社。

贾伟明,1987,《数学方法在考古学研究中应用的探讨》,《考古学文化论集》(1),文物出版社。

韩康信,潘其风, 1985,《安阳殷墟中小墓人骨的研究》、《殷墟祭祀坑人头骨的种系》,《安阳殷墟头骨研究》,文物出版社。

华觉明,1999,《中国古代金属技术——铜和铁造就的文明》,大象出版社。

湖北省江陵博物馆,1984,《江陵雨台山楚墓》,文物出版社。

黄其煦,1988,《安阳殷墟中小墓中人骨的对应分析》,《考古》1988年4期。

黄蕴平,1996,《动物化石》,《南京人化石地点(1993—1994)》,文物出版社。

李非,李水城,水涛,1996,《葫芦河流域的古文化和古环境》,《考古》1996年9期。

李国霞,赵维娟,李融武等,2002,《古耀州瓷胎起源的模糊聚类分析》,《科学通报》Vol 47, No23。

李晓岑等,2000,《中国铅同位素考古》,云南科技出版社。

林沄,1980,《中国东北系铜剑初论》,《考古学报》1980 年 2 期。

卢淑华,1989,《社会统计学》,北京大学出版社。

刘茂,1989,《浅谈考古学研究中的定量分析问题》,《史前研究》闭刊专辑。

楼世博,孙章,陈化成,1983,《模糊数学》,科学出版社。

罗立强,郭常霖,吉昂等,1997,《人工神经网络与分析测试技术的研究和发展》,《岩矿测试》16 卷 4 期。

米同乐,戴书田,1998,《有胡铜戈的回归断代》,《河北省考古文集》,东方出版社。

苗建民,汪安,陆寿龄,1993,《古陶瓷中痕量元素的模糊聚类分析》,《科学通报》1993 年 4 期。

苗建民,王时伟,2005,《紫金城清代剥釉琉璃瓦件施釉重烧的研究》,《故宫学刊》2005 年第 1 辑。

闫渭清,1991,《试论"概率分析"在考古学中的运用》,《西北史地》1991 年 1 期。

裴安平,李科威,1991,《雨台山楚墓 CASA 年代序列分析与相关问题讨论》,《考古》1991 年 5 期。

滕铭予,2000,《多变量分析及其在考古学研究中的应用》,《考古学集刊》(13),文物出版社。

滕铭予,2004,《数学方法在考古类型学研究中的实践和思考》,《边疆考古研究》第 2 辑,科学出版社。

王建平,陈铁梅,2003,《广东博罗先秦陶瓷的 INAA 研究》,《核技术》26 卷(6)。

王志俊,1980,《关中地区仰韶文化刻划符号综述》,《考古与文物》1980 年 3 期。

吴十洲,2001,《两周墓葬青铜容器随葬组合定量分析》,《考古》2001 年 8 期。

西安半坡博物馆,渭南文化馆,1978,《陕西渭南史家新石器时代遗址》,《考古》1978 年 1 期。

谢衷洁,2004,《普通统计学》,北京大学出版社。

阳含熙,卢泽愚,1981,《植物生态学的数量分类方法》,科学出版社。

杨希枚,1985,《河南安阳殷墟墓葬中人体骨骼的整理和研究》,《安阳殷墟头骨研究》,文物出版社。

伊竺,1985,《关于元君庙,史家乡仰韶墓地的讨论》,《考古》1985 年 9 期。

张宗培,1981,《史家墓地研究》,《考古学报》1981 年 2 期。

朱乃诚,1984,《概率分析方法在考古学中的初步应用——以渭南史家墓地为分析对象》,《史前研究》1984 年 1 期。

Aldenderfer M.S(editor), 1987, *Quantitative Research in Archaeology: Progress and Prospects*, SAGE Publications Inc..

Baxter M.J., 1994, *Exploratory Multivariate Analysis in Archaeology*, Edinburgh University Press.

Brainerd G.W., 1951, "The place of chronological ordering in archaeological analysis", *American Antiquity*, 16.

Binford L. R. and Binford S. R., 1966, "A preliminary analysis of functional variability in the Mousterian of Levallois facies", *American Anthropology*, 68.

Champion T., P. Cuming and S. J. Shannan, 1996, *Planning for the Past* Vol. 3: Decision-making and field methods in Archaeological Evaluation, London: English Heritage and University of Southampton.

Clarke D. L., 1979,《考古学纯洁性的丧失》("Archaeology: The loss of innocence"的译文),《考古学文化论集》(2),1989,文物出版社。

Chen Tiemei, Chang Yingung, 1991, "Palaeolithic chronology and possible coexistence of *Homo erectus* and *Homo sapiens* in China", *Wored Archaeology* 23(2).

Dempsey P. and Baumhoft M., 1963, "The statistical use of artificial distribution to establish chronoogical sequensce", *American Antiquity*, 28.

Drennan R. D., 1996, *Statistics for Archaeologists-A Commonsense Approach*, Plenum Press.

Fletcher M. and Lock G. R., 2001, *Digging Numbers - Elementary Statistics for Archaeologists*, published by Oxford University School of Archaeology.

Joseph F. and Hair Jr., 1979, *Multivariate Data Analysis*, Petroleum Publishing Company.

Legge A. L. and Rowley-Conwy P., 1986, "New radiocarbon dates for early sheep at Tell Abu Huryra, Syria", in *Archaeologic Result from Accelerator Dating*, Oxford University Press.

Ma QL, Yan AX, Hu ZhD, et al., 1999, "Principal component analysis and artificial neural networks applied to the classification of Chinese pottery of neolithic age", *Analytica Chimica Acta*, 20225 (1999).

Robinson W. S., 1951, "A method for archaeological ordering archaeological deposits", *American Antiquity*, 16.

Shennan S., 1997, *Quantifying Archaeology*, second edition, Edinburgh University Press.

Spatz C. & Johnson J. O., 1989, *Basic Statistics-Tales of Distributions*, fourth Edition, Brooks/Cole Publishing Company.

Tukey W., 1977, "Exploratory Data Analysis", Reading MA Addison-Wesley.

Wainwright G. J., 1979, Mount Pleasant, "Dorset: Excavations 1970 - 1971", Society of Antiquaries Research Report 37, London: Thames & Hudson.

Wright R., 1989, "Doing multivariate archaeology and prehistory: Handling large data sets with MV-Arch", Sedney: Department of Anthropology, University of Sydney.

Yuan Jing, Liang Zhonghe, Wu Yun, et al., 2002, "Shell mound in the Jiaodong Peninsula: a study in environmental archaeology", *Journal of East Asia Archaeology*, Vol 4, 1 - 4.

Г. Л. Федов Давыдов, 1987, Статистические Методы в Археологии, Издательство, *Высшая школа*, Москва.

附录一 习 题

上篇 考古研究中的基础统计学

第三章

1. 列式计算下列一组数据的平均值,离差平方和,样本标准差和总体标准差,中数,四分位数和四分位差。能否求众数,为什么?(请手工计算)

55,53,57,51,52,62,55,55,56,54。

2. 有一批青铜剑,其长度分别为120,121,130,125,126,128,126,135,125,86,82,94,87,89,85,89,126,124,82,86,125,87厘米。请画出直方图,问求它们的平均长度和长度标准差有无意义?为什么?这组数据应如何处理?

3. 从Pine Ridge Cave (PRC) 和 Willow Flats Site (WFS) 两地点采集到48件燧石质刮削器,并测量了它们的长度(单位为毫米)。石质可分为Chert (C) 和 Flint (F) 两种。48件刮削器的出土地点,石质材料和长度统计如下表。请分别以地点和石质分类,画出两张背对背的茎叶图(或相应的直方图),并加以讨论。(引自匹兹堡大学教材 Drennan1996)

地点	PRC	PRC	PRC	PRC	PRC	PRC	PRC	PRC	PRC	PRC
材料	C	C	F	C	F	C	C	C	C	C
长度	25.8	6.3	44.6	21.3	25.7	20.6	22.2	10.5	18.9	25.9
地点	PRC	PRC	PRC	PRC	PRC	PRC	PRC			WFS
材料	C	C	C	F	C	C	F			C
长度	23.8	22.0	10.6	33.2	16.8	21.8	48.3			15.8
地点	WFS	WFS	WFS	WFS	WFS	WFS	WFS	WFS	WFS	WFS
材料	F	F	F	C	F	F	C	C	F	F
长度	39.4	43.5	39.8	16.3	40.5	91.7	21.7	17.9	29.3	39.1
地点	WFS	WFS	WFS	WFS	WFS	WFS	WFS	WFS	WFS	WFS
材料	F	F	C	C	F	F	F	C	C	F
长度	42.5	49.6	13.7	19.1	40.6	49.1	41.7	15.2	21.2	30.2
地点	WFS	WFS	WFS	WFS	WFS	WFS	WFS	WFS	WFS	WFS
材料	F	C	F	F	F	F	C	C	F	F
长度	40.0	20.2	31.9	42.3	47.2	30.5	10.6	23.1	44.1	45.8

4. 测量了28件周代青铜剑中锡的百分含量(见下表),请画出直方图和茎叶图,计算平均值、标准差、标准误、中值、上下四分位数和四分位差。

11.5 12.3 12.4 12.4 13.5 14.3 14.3 14.5 14.6 14.6 14.7 15.0

15.7 15.8 15.8 15.9 16.2 17.1 17.5 17.6 17.7 18.4 18.6 18.8
19.0 19.7 19.8 20.5

5. 下表的数据引自华觉明(1999)的表 7-11 和表 7-12。统计了从西周晚期和东周青铜剑的铜锡铅元素组成。表的右面为中原的数据,左面为辽西的数据。请分析比较两地剑中锡的平均含量,对锡含量稳定性的控制(首先请注意辽西剑锡含量的分布)。也请比较两地铸剑使用铅的情况。如有可能,请尝试作考古学的解释。

中原地区					辽 西				
编号	出土地点	铜	锡	铅	编号	出土地点	铜	锡	铅
1	洛阳	88.22	11.76		1	建平	95	1.5	3.5
2	洛阳	81.05	16.17		2	朝阳	88	5	7
3	洛阳	84.4	12.32	1.96	3	朝阳	91	3	6
4	洛阳	83.3	14.61	0.72	4	朝阳	84	3	13
5	洛阳	78.96	19.78	0.29	5	朝阳	84	3	13
6	洛阳	77.62	20.5	0.83	6	朝阳	84	4	12
7	洛阳	74.27	14.57	7.54	7	朝阳	77	4	19
8	洛阳	74	12.4	10.37	8	朝阳	87	2	11
9	洛阳	73.85	17.48	6.84	9	朝阳	77	4	19
10	洛阳	73.72	19.01	6.97	10	朝阳	75	2	23
11	洛阳	73.7	12.36	12.46	11	朝阳	84	3	13
12	洛阳	72.26	19.65	7.14	12	建平	65	14	21
13	洛阳	72.15	15.78	10.71	13	建平	75	8	17
14	洛阳	72.12	15.92	11.24	14	朝阳	78	8	14
15	洛阳	71.59	17.14	10.4	15	建平	73.2	14.2	12.6
16	洪洞	73.39	17.7	6.67	16	朝阳	61.7	21.8	16.4
17	江陵	80.3	18.8	0.4	17	朝阳	60	20	20
18	沈阳	72.43	13.52	6.84	18	朝阳	62.8	23	13.1
19	罗定		11.5	0.13	19	朝阳	78	12	10
20	罗定		17.6	1.1	20	朝阳	72	13	15
21	江西		18.6		21	朝阳	78	5	17
22	涪陵	82.21	14.67	1.28	22	朝阳	67	20	13
23	云南	80.52	15.67		23	朝阳	76	5	18
24	涪陵	84	14.29	1.51	24	朝阳	62	19	19
25	云南	83.4	14.45	1.33	25	朝阳	79	13	8
26	长沙	78.38	14.98	1.48	26	朝阳	77	7	16
27	长沙	71.61	14.33	10.2	27	朝阳	73	8	19
28	长沙	73.79	18.42	1.03	28	朝阳	70.5	13	16.5
29	龙门		15.76	2	29	朝阳	74	13	13
30		78.48	19.88	0.25	30	朝阳	83	25	14
31		79.13	19.35	0.19	31	朝阳	73	11	16
32		80.33	17.73	0.25	32	朝阳	79	13	8
33		84.58	11.76	2.13	33	朝阳	78	8	14

续表

中原地区					辽　西				
编号	出土地点	铜	锡	铅	编号	出土地点	铜	锡	铅
34	河南	76	16.13		34	北票	85	0.01	14.1
35		62.59	14.54	15.44	35	北票	84	3	13
36		82.32	16.44		36	北票	84	11	5
37		81.06	15.8	0.12	37	朝阳	74.5	13.6	11.9
38		81.75	15.42	0.1					
39		81.1	17.13	0.3					
40		80.35	16.56	0.27					
41		74.86	18.4	4.81					
42		75.1	18.6	4.87					
43		73.14	19.18	6.39					
44		68.09	14.29	8.39					
45		73.34	19.84						
46		69.31	12.55	7.92					
47		66.6	14.13	1.32					
48		71.93	16.19	10.83					

第四章

1. 抽扑克牌，每次抽一张，抽完后放回，混匀，再抽第二张，问：

(1) 抽三张牌，花色一样的概率是多少？

(2) 抽三张牌，不计花色，三张牌顺序相连的概率是多少(认为 A 同时连接 2 和 K)

(3) 抽三张牌，数值一样的概率是多少？

(4) 抽二张牌，其和大于 5 的概率是多少？

2. 如果抽出的牌，不再放回，上题的计算结果如何？

3. 对某居民小区进行了调查，统计有 80% 的居民订阅报纸，45% 的居民订阅杂志，30% 的居民同时订阅报纸和杂志。请计算只订阅报纸的居民的百分比和不订阅任何报纸杂志的居民比例。

4. 据统计某城市居民活到 60 岁的概率是 80%，活到 70 岁的概率是 40%，问现年 60 岁的人活到 70 岁的概率是多少？(引自卢淑华(1989))

5. 某种产品由甲乙两工厂提供，已知甲厂提供 95%，其次品率为 2%，乙厂提供 5%，其次品率为 20%。现发现一件次品，问它是甲厂生产的概率多大。

6. 有一副扑克牌，由等数量的 A = 1,2,3,4 组成。
(1)随机抽一张，得一个值 X_1。求 X_1 的理论分布应如何？
(2) 随机抽一张，得一个值 X_1，抽完后放回，混匀，再抽第二张，求两次平均值 X_2 的理论分布。
(3) 按上述规则，依次抽三张，问三张牌的平均值 X_3 的理论分布应如何？
(4)怎样认识上述分布的变化？

7. 发现有 6 个墓，请计算墓主人性别分别为 0,1,2,3,4 个男性的概率(假设男女性比正常)。

8. 已知北大男同学平均身高为 171cm，标准差为 4cm，假设身高服从正态分布，问：
(1) 身高大于 179cm 的人的百分比？
(2) 要选 20% 中等身高的人，请定这批人身高的上下限，即身高在此区间的人占总数的 20%。
(3) 随机找一个同学，其身高在 1.72—1.75cm 间的概率是多少？
(4) 有一男生，其身高属最高的 5% 以内，问其身高应不低于多少？
(5) 招考飞行员，身高要求在 1.68—1.78cm 之间，问有多少比例的人不能报名？

9. 第 8 题的 a, b，请在标准差为 2cm 的条件下求解。比较这二种情况的结果，并作讨论。

10. 计算第三章第 1 题中各数据的 Z 分量值，并验证 Z 的平均值和标准差是否接近 0 和 1。

第五章

1. 测量了一组 $k = 49$ 把同类型青铜剑的长度(已知它们服从正态分布)，分别用 X_1，X_2，X_3，$\cdots X_{49}$ 表示(单位为 cm)，已求出这批青铜剑长度的平均值和标准差为 50cm 和 6cm。问(1)计算任意抽取一剑其长度在 44—56cm 间的概率，(2)求该组数据的标准误并给出该类型剑平均长度 50% 置信度的区间估计，(3)如果我们希望将对该类型剑平均长度 50% 置信度的区间估计的宽度缩短一半，问至少需要测量多少把剑。

2. 将第三章第 4 题的数据看成大样本 ($n = 28$)，请给出这类青铜剑平均含锡量 68.3% 和 95% 置信度的区间估计。

3. 如果将上题作为小样本处理，这类青铜剑平均含锡量 68.3% 和 95% 置信度的区间估计是多少，与上题的差别有多大。

4. 假设某文化类型的聚落面积服从正态分布，已测得五个该类聚落的面积分别为50，60,65,55和70平方里，请分别给出0.1和0.05的显著性水平下该类聚落平均面积的区间估计并作比较。

5. 以下一些样本来自平均值 $\mu = 29.00$ 的总体，求它们的 t 值。

(a) $\bar{X} = 29.60 \quad s = 2.50 \quad n = 29$

(b) $\bar{X} = 29.60 \quad s = 0.15 \quad n = 29$

(c) $\bar{X} = 29.60 \quad s = 0.15 \quad n = 3$

(d) $\bar{X} = 25.00 \quad s = 0.15 \quad n = 29$

请观察 t 值的大小依赖于哪些因素(请在a,b间、b,c间和b,d间比较)。

第六、七章

1. 根据第三章第5题中原地区48把周代青铜剑实测的锡含量，检验《六齐说》关于周代青铜剑是按照锡含量25%铸造的说法是否正确。

2. 对某遗址地面和半地下房屋面积S统计如下：

地面　统计房屋数 $n = 50$　平均面积 $S = 7\text{m}^2$，　方差 $= 5\text{m}^2$，

半地下　$n = 80$　平均面积 $S = 6.6\text{m}^2$，　方差 $= 6\text{m}^2$，

计算两类房屋面积平均值之差的0.05显著性水平的区间估计，并在一定的显著性水平下作出推论，问这两类房屋的平均面积有没有显著差异。

3. 调查某城市男女平均寿命，结果如下：

性别	被调查人数	平均寿命	标准差
男	150	76	11
女	100	80.5	12

问：(1)该城市男女的平均寿命有没有差异？

(2) 如果标准差分别为18与20，重复上面的判断。

4. Pittsburgh大学的R.D. Drennan完成了对Ollantaytambo的考古发掘，找到了36件燧石工具，其外表颜色和锆含量的测定如下：

黑	灰	黑	灰	灰	灰	灰	黑	黑	灰
137.6	133.3	137.3	137.1	138.9	138.5	137.0	138.2	138.4	135.8
黑	黑	黑	黑	灰	灰	黑	灰	灰	灰
137.4	140.0	136.4	138.8	136.8	136.3	135.1	132.9	136.2	139.7

续表

黑	灰	灰	灰	灰	黑	黑	灰	黑	灰
139.1	139.2	132.6	134.3	138.6	138.6	139.0	131.5	142.5	137.4
黑	灰	黑	灰	黑	黑				
141.7	136.0	136.9	135.0	140.3	135.7				

从外表颜色分析,不同颜色的燧石工具似乎有不同的原材料来源，但 Drennan 希望通过其化学元素组成来验证这个判断,表中列出了燧石工具的颜色和锆含量的测定结果(用 ppm 作单位表示)。

(1) 请对不同颜色的燧石工具画出背对背的茎叶图作初步验证。

(2) 分别计算两种颜色燧石工具总体的锆含量平均值 0.05 显著性水平的区间估计。

(3) 请在 $\alpha = 0.05$ 的显著性水平下检验两种不同颜色燧石工具的平均锆含量是否有差别,请讨论燧石工具的锆含量分析能否佐证工具外表色泽对原材料产地来源的判断。

(4) 请同时画出两种颜色燧石工具按锆含量分布的箱点图,并作讨论。

5. 抽样调查测量了甲乙两地部分(8 个和 10 个)聚落的面积,记录如下(为便于计算，表的最后 2 列列出 ΣX 和 ΣX^2)。

											ΣX	ΣX^2
甲	64	67	72	57	60	71	62	67			520	33992
乙	69	74	64	76	74	70	71	62	72	68	700	49178

(1)根据样本的数据分别给出甲乙两地聚落平均面积 95%置信度的区间估计。

(2)分别在 0.05 和 0.02 的显著性水平上,检验两地聚落的平均面积有无明显差别，如有,请对差别的大小作出区间估计。

(3)如可能请说明本题中检验平均面积有无明显差别时,需要什么前提条件,怎样检验或考察前提条件是否成立,如前提条件不满足,有什么其他方法来检验。

6. Cottonwood River Valley 经典期早晚三段的部分尖状器的重量统计如下：

	早期	中期	晚期	全部
统计数量 n	58	42	27	127
平均值 (g)	53.67	60.45	41.56	53.34
样本标准差 (g)	14.67	12.15	8.76	14.42
标准误 (g)	1.93	1.88	1.69	1.28
样本方差 (g^2)	215.21	147.62	76.74	207.94

有人试图由尖状器重量随时间的变化来探讨狩猎大,小动物的比例随时间的变化。请用一元方差分析(ANOVA)处理上述数据并进行有关讨论。

7. 同时期同地区三种类型聚落的面积统计如下，问三种类型聚落的平均面积有无显著差别。

类型 A 23,18,31,14,22,28,20,22

B 17,20,22,19,21,14,25

C 16,11,13,10,12,10

已算出 对 A $\sum x = 178$, $\sum x^2 = 4162$, 对 B $\sum x = 138$, $\sum x^2 = 2796$,

对 C $\sum x = 72$, $\sum x^2 = 890$,

8. 分别用 X 荧光分析 XRF 和中子活化分析 INAA 方法测量了 28 片隋唐时期洪州窑瓷片中氧化钡含量。请在一定的显著性水平下检验这两种方法的测量数据之间是否存在系统误差，如果存在差别，请估计差别的数值。还请用符号检验方法检验是否存在系统误差。测量数据见下表：

XRF	475	417	572	577	502	473	514	563	557	567	362	615	604	587
INAA	479	378	552	533	532	456	487	500	537	521	377	606	585	574

XRF	565	653	704	605	595	488	544	576	515	612	503	517	581	516
INAA	592	590	607	621	648	539	530	612	544	582	518	642	526	535

9. 请对本章第 5 题的数据作非参数的秩和检验，在一定的显著性水平下判断，甲乙两地聚落的平均面积有无显著差别。

第八章

1. 设有一墓地，发现有 10 具成年人骨，请分别在 $\alpha = 0.1$ 和 0.01 的显著性水平下讨论，男性人骨要达到多少具时，才能认为墓地所属氏族的男女性比不正常（假设人骨性比能代表墓地所属氏族的男女性比）（提示：先计算 10 男，9 男 1 女……的概率，再进行讨论）。如果成年人骨为 1000 具应如何解题。

2. 某旧石器时代遗址随机地面采集了 200 件石器，其中 70 件为燧石工具，请分别以 95% 和 99% 的置信度估计该遗址燧石工具所占百分比的范围。如果要求对燧石工具所占百分比的估计精（密）度达到 ±1%，置信度为 95%，问至少应采集多少片石器。请用文字准确地表述本题的解。

3. 对两个新石器早期遗址作调查，分别统计了 100 件动物个体骨骼，人工饲养动物的比例分别占 45% 和 60%，请在一定的显著性水平上讨论该两遗址家畜饲养的发展水平有无差异。如有差异，请给出差异的区间估计。

4. 在两个墓地分别统计部分墓葬,每地各统计了 100 个墓葬,发现甲墓地有仰身葬 45 个,乙墓地有仰身葬 63 个。问两个墓地仰身葬的比例有无显著差别?($\alpha = 0.05$)

5. 我们知道当 $n > 30$ 时,二项式分布十分接近正态分布,请用 U 检验以下墓地所属氏族的男女性比是否正常(假设人骨性比能代表墓地所属氏族的男女性比)。

(1) 永昌,鸳鸯池 半山马厂墓地 男 29 具, 女 24 具(均为成年)
(2) 华县,元君庙仰韶墓地 男 85, 女 61
(3) 兖州,王因大汶口墓地 男 547, 女 233

6. 某化石动物群鉴定了 50 个个体,未见到披毛犀,你有多少的置信度断言,该动物群中披毛犀的百分比低于 1%。如果你希望上述判断的置信度提高到 95%,至少应鉴定多少个个体。

第九、十、十一章

1. 下面是某地旧石器时代手斧的重量和上面打击痕数目的统计。请先画散点图(应怎样选自变量),再判断手斧重量和打击痕数目间是否相关,相关强度如何,有什么考古意义?请写出线性回归方程。

打击痕数	18	19	33	28	24	36	45	56	47
重量(克)	210	300	195	285	410	375	295	415	500
打击痕数	37	72	57	53	46	78	68	63	82
重量(克)	620	510	565	650	740	690	710	840	900

2. 已知某地区存在 A,B,C 三种不同生态环境的土地,并各占面积为 39, 83 和 14 平方公里。考古调查在这三种土地上依次发现有 19,12 和 7 个同时期聚落遗址。请判断古人对聚落地点的选择是否考虑地点的生态环境。

3. 表中给出从两个遗址地面随机采集到的甲乙丙三种风格陶片的数目:

	A	B
甲	162	40
乙	49	43
丙	57	49

请判断这两个遗址甲乙丙三种风格陶片的相对比例在一定的显著性水平上有无差异,即“遗址”和“陶片风格组成”这两个变量之间有没有关联,并讨论关联强度。

4. 调查了 140 件某种类型的瓶子，按其瓶口和瓶颈形式，有无纹饰统计如下：

口沿形式	瓶颈形式	有无纹饰	数量	
X	A	有	16	
X	B	有	9	
X	A	无	14	
X	B	无	32	
Y	A	有	7	
Y	B	有	14	
Y	A	无	30	
Y	B	无	18	
			总计 140	

请分析口沿和瓶颈之间有无关联，关联强度有多大，并分析当引入第三变量纹饰后，口沿和瓶颈之间关联强度的变化。

5. 下表是对某墓地 10 座墓葬分别按它们的墓穴大小和随葬品的质和量排列的次序，请判断墓穴大小和随葬品的质量之间是否相关。

墓号	A	B	C	D	E	F	G	H	I	J
墓穴大小次序	1	2	3	4	5	6	7	8	9	10
随葬品质量次序	3	3	5	6	1	10	3	7	8	9

6. 下面两表分别列出朱乃诚—陈铁梅和朱乃诚—刘茂对渭南史家墓地墓葬分期的比较（见第十七章）。请分别计算两表的 GAMMA、Kendall's τ_b 和 τ_c 等级相关系数，并作讨论（不考虑朱乃诚所定 IV 期的墓葬）。

朱乃诚 \ 陈铁梅	I	II	III	IV
I	3	0	0	0
II	1	2	0	0
III	5	5	5	0
V	0	3	1	1
VI	0	0	1	5

朱乃诚 \ 刘茂	I	II	III
I	3	0	0
II	1	0	2
III	7	6	2
V	2	0	3
VI	1	2	7

下　篇　多元统计分析

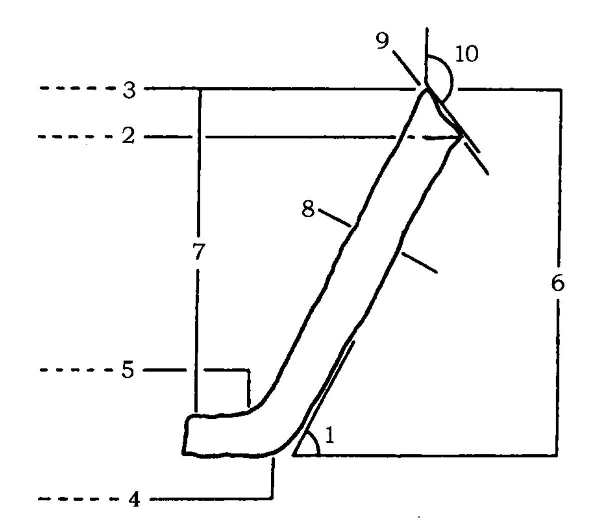
下面两张用于多元统计分析的数据表均引自 Shennan(1997),读者可以直接使用这两张表的数据,或使用表中线性尺度数据的比值,也可以使用本书表 14-7 和表 15-28 的数据,作多种聚类分析、K-均值分类和主成分分析,并对分析结果作比较。

1. 两河流域乌鲁克时期 42 只斜边碗的 10 项测量数据表,各测量项目的意义见附录图-1。

编号	底角	口外径	内口径	底外径	内底径	高	里高	壁厚	口厚	口角
1	58	160	150	80	70	73	65	108	145	128
2	57	140	130	70	65	67	62	94	111	137
3	55	175	155	70	70	71	61	107	110	137
4	58	180	170	70	65	84	80	106	121	154
5	62	195	180	80	70	86	72	108	135	150
6	60	165	160	70	65	85	78	111	130	159
7	53	180	170	80	65	85	75	120	123	148
8	68	130	120	60	50	71	65	108	104	150
9	48	150	140	70	60	70	55	133	129	165
10	58	200	190	80	75	96	84	159	141	147
11	47	210	200	85	75	79	74	114	135	163
12	60	160	150	80	70	87	80	110	121	136
13	55	180	170	80	80	88	83	109	118	160
14	65	190	152	80	75	91	79	132	169	150
15	63	190	170	75	70	89	85	137	129	155
16	67	220	210	80	75	118	105	145	138	170
17	44	170	150	80	70	58	44	103	123	154
18	63	185	170	75	80	80	74	117	139	148
19	52	160	150	60	55	75	69	109	126	148
20	62	215	200	90	85	97	81	138	128	133
21	41	175	160	65	60	70	62	110	137	151
22	47	190	170	75	80	69	58	120	129	148
23	50	185	160	70	65	94	80	126	143	152
24	55	195	180	70	65	85	80	130	129	151
25	49	195	180	70	65	77	69	124	102	148
26	58	140	120	65	60	66	54	113	143	130
27	62	170	160	65	60	90	70	94	131	137
28	55	136	120	70	65	73	64	109	102	136
29	53	170	160	70	65	78	64	123	124	135
30	60	175	160	70	60	83	70	112	142	155
31	52	140	120	70	65	73	62	116	126	145

续表

编号	底角	口外径	内口径	底外径	内底径	高	里高	壁厚	口厚	口角
32	59	150	140	75	70	88	76	101	126	135
33	61	140	130	70	60	92	85	116	103	152
34	56	145	130	65	60	72	65	125	134	136
35	60	175	160	75	65	93	78	111	160	130
36	53	165	160	70	60	74	65	111	62	160
37	49	165	150	80	75	75	62	129	147	154
38	60	160	140	70	65	78	66	114	146	143
39	59	170	160	70	60	91	77	138	119	146
40	57	165	160	80	63	77	60	91	124	170
41	55	170	160	80	65	70	66	140	121	149



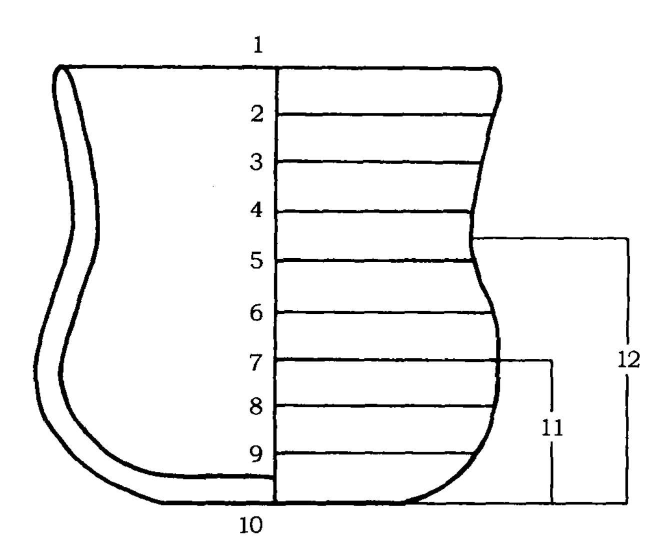
附录图-1　两河流域乌鲁克时期斜边碗10项测量项目的示意图。1底角,2口沿外径,3口沿内径,4底外径,5底内径,6高,7里高,8边壁厚,9口沿壁厚,10口沿角。

2. 欧洲中部地区新石器晚期22件陶罐的12项描述其形状的测量数据表(见附录图-2)

	1	2	3	4	5	6	7	8	9	10	11	12
1	60.36	55.86	51.35	48.65	50.45	53.15	54.05	50.45	42.34	27.93	37.84	65.77
2	41.28	37.61	35.78	36.70	39.45	43.12	42.20	38.53	33.03	25.69	38.53	77.06
3	40.96	38.55	37.35	37.35	48.19	53.01	54.22	50.60	43.37	21.69	33.73	68.67
4	34.88	34.88	38.37	40.75	50.00	56.98	59.30	55.81	47.67	33.72	34.88	62.79
5	54.84	50.54	47.31	48.39	53.76	59.14	62.37	58.06	46.24	31.18	34.41	75.27
6	47.62	41.90	39.05	40.00	41.90	44.76	45.71	42.86	36.19	20.00	36.19	70.48
7	38.40	34.40	32.00	32.00	33.60	36.80	39.20	38.40	32.00	16.80	30.40	71.20
8	40.00	36.47	34.12	35.29	36.47	41.18	44.71	42.35	36.47	17.65	36.47	72.94
9	48.24	44.71	38.82	35.29	31.76	38.82	40.00	35.29	25.88	15.29	37.65	51.76
10	37.50	31.94	30.56	31.25	34.72	38.89	42.36	40.28	34.03	24.31	31.25	64.58
11	32.18	27.59	28.74	49.43	52.87	50.57	47.13	41.38	33.33	22.99	55.17	81.61

续表

	1	2	3	4	5	6	7	8	9	10	11	12
12	32.86	34.29	37.14	42.86	48.57	51.43	50.00	44.29	34.29	8.57	41.43	84.29
13	50.75	47.76	47.76	64.18	70.15	70.15	64.18	56.72	41.79	20.90	49.25	79.10
14	35.71	34.52	35.71	39.29	44.05	46.43	45.24	39.29	30.95	20.24	41.67	66.67
15	35.29	34.31	33.33	36.27	44.12	49.02	50.98	49.02	41.18	20.59	32.35	70.59
16	37.33	36.00	36.00	45.33	54.67	61.33	62.67	60.00	48.00	21.33	37.33	78.67
17	44.00	42.67	41.33	41.33	50.67	56.00	57.33	54.67	42.67	21.33	36.00	69.33
18	51.39	45.83	38.89	37.50	37.50	40.28	44.44	45.83	37.50	22.22	26.39	59.72
19	46.74	43.48	40.22	41.30	44.57	48.91	52.17	46.74	38.04	22.83	32.61	63.04
20	32.17	32.17	31.30	33.04	39.13	43.48	44.35	42.61	35.65	21.74	34.78	62.61
21	50.53	48.42	48.42	54.74	60.00	62.11	62.11	58.95	48.42	27.37	36.84	73.68
22	66.15	64.42	56.92	52.31	52.31	55.38	55.38	53.85	46.15	33.85	41.54	56.92



附录图-2 中欧新石器晚期陶罐形状的示意图(12个测量指标)

附录二　利用 Excel 软件计算几个常用统计函数的数值

（一）计算排列数和组合数的函数

$$\text{PERMUT}(n,m) = P_n^m = \frac{n!}{(n-m)!} \tag{附 -1}$$

$$\text{COMBIN}(n,m) = C_n^m = \frac{P_n^m}{m!} = \frac{n!}{(n-m)! \cdot m!} \tag{附 -2}$$

公式中 n 和 m 均为正整数。n 表示对象的个数，m 表示被选对象的个数，$m \leqslant n$。例如 PERMUT(4,3) = 24，COMBIN(4,3) = 4。

（二）二项式分布函数

假设单次贝努里试验成功的概率为 p，失败的概率为 q，且 $p + q = 1$，则 n 次实验中成功 m 次的概率服从二项式分布 $C_n^m p^m q^{(n-m)}$。EXCEL 软件的相应函数为

$$\text{BINOMDIST}(m,\ n,p,FALSE) = P\{\xi = m\} = C_n^m p^m q^{(n-m)} \tag{附 -3a}$$

$$\text{BINOMDIST}(m,\ n,p,TRUE) = P\{\xi \leqslant m\} = \sum_{\zeta=0}^{m} C_m^{\zeta} p^{\zeta} q^{(n-\zeta)} \tag{附 -3b}$$

BINOMDIST 函数的自变量为 m，有 2 个参数 n，p 和 1 个开关参数，开关参数选择“*FALSE*”或“*TRUE*”决定函数返回微分概率或积分概率。例如 BINOMDIST(2, 4, 0.5, *FALSE*) = 0.375, BINOMDIST(2, 4, 0.5, *TURE*) = 0.6875。

（三）正态分布函数

正态分布函数 $N(x,\mu,\sigma)$ 是包含 2 个参数 μ 和 σ 的函数，μ 和 σ^2 分别代表正态函数的数学期望值和方差。正态分布函数的分析形式如下

$$N(x,\mu,\sigma) = f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

EXCEL 软件的相应函数为

$$\text{NORMDIST}(x,\ \mu,\ \sigma,\ FALSE) \tag{附 -4a}$$

返回函数的数值 $f(x)$，即返回概率密度值。和

$$\text{NORMDIST}(x,\ \mu,\ \sigma,\ TRUE) \tag{附 -4b}$$

返回函数的积分值 $\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$，即返回累积概率 $P\{-\infty < \xi < x\}$。NORMDIST 函数中需对开关变量赋值，“FALSE”或“TRUE”。

EXCEL 软件还提供正态分布函数的反函数，即已知累积概率 $P\{-\infty < \xi < x\}$，计算 x 值。

$$\text{NORMINV}(\text{累积概率值 } P,\mu,\ \sigma) \tag{附 -5}$$

返回对应的 x 值。

EXCEL 软件的内部函数中包含 $\mu = 0$，$\sigma = 1$ 的标准型正态分布函数及其反函数，它们是

$$\text{NORMSDIST}(x) \qquad (附\text{-}6)$$

返回累积概率 $P\{-\infty < \xi < x\}$。(对于标准型正态分布函数,EXCEL 软件不能计算概率密度值,相当于 NORMDIST 函数中的开关值永远是“*TRUE*”)。

$$\text{NORMSINV}(累积概率值\ P) \qquad (附\text{-}7)$$

返回对应的 x 值。例如 NORMSDIST(0) = 0.5, NORMSINV(0.975) = 1.96

(四) t 分布函数

t 分布函数的分析表达式 $f(x)$ 比较复杂,它只有一个参数 - 自由度 df。EXCELl 软件提供的函数为,

$$\text{TDIST}(x, df, 1) \qquad (附\text{-}8a)$$

$$\text{TDIST}(x, df, 2) \qquad (附\text{-}8b)$$

式中的“1”和“2”为开关值。这两个式子分别返回单边或双边的累积概率值 $P\{-\infty < \xi < x\}$ 或 $(1 - P\{-x < \xi < x\})$。t 分布的反函数为

$$\text{TINV}(双边的累积概率值, df) \qquad (附\text{-}9)$$

返回对应的 x 值。

例如 TDIST(1.96,6,2) = 0.0977,TDIST(1,96,6,1) = 0.0488, TINV(0.977,6) = 1.96。

(五) χ^2 分布函数

χ^2 分布函数与 t 分布函数相似,只有一个参数 - 自由度 df。相应的 EXCEL 内部函数有

$$\text{CHIDIST}(x, df) \qquad (附\text{-}10)$$

返回 χ^2 大于 x 的尾部累积概率(当函数的自由度为 df 时) $P\{\chi^2 > x\} = \int_x^\infty f(\chi^2)\,d\chi^2$。其反函数为

$$\text{CHIINV}(尾部累积概率值\ P, df) \qquad (附\text{-}11)$$

返回对应于尾部累积概率为 P 时的 x 值。例如 CHIDIST(10,6) = 0.125,和 CHIINV(0.125, 6) = 9.991。

(六) F 分布函数

F 分布函数是双参数函数,两个参数分别为第一自由度(分子自由度) $df1$ 和第二自由度(分母自由度) $df2$。相应的 EXCEL 内部函数是

$$\text{FDIST}(x, df1, df2) \qquad (附\text{-}12)$$

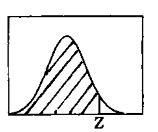
返回 F 大于 x 的尾部累积概率(当函数的自由度为 $df1$ 和 $df2$ 时) $P\{F > x\} = \int_x^\infty f(F)\,dF$。相应的反函数为

$$\text{FINV}(尾部累积概率值\ P, df1, df2) \qquad (附\text{-}12)$$

返回对应于尾部累积概率为 P 时的 x 值。例如 FDIST(10,2,6) = 0.0123 和 FINV(0.0123, 2,6) = 9.996。

附录三　标准型正态分布临界值表

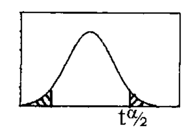
$$\Phi(Z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Z} e^{-\frac{t^2}{2}} dt$$



Z	Φ(Z)	Z	Φ(Z)	Z	Φ(Z)	Z	Φ(Z)
0	0.5000	0.62	0.7324	1.24	0.8925	1.86	0.9686
0.02	0.5080	0.64	0.7389	1.26	0.8962	1.88	0.9699
0.04	0.5160	0.66	0.7454	1.28	0.8997	1.9	0.9713
0.06	0.5239	0.68	0.7517	1.3	0.9032	1.92	0.9726
0.08	0.5319	0.7	0.7580	1.32	0.9066	1.94	0.9738
0.1	0.5398	0.72	0.7642	1.34	0.9099	1.96	0.9750
0.12	0.5478	0.74	0.7704	1.36	0.9131	1.98	0.9761
0.14	0.5557	0.76	0.7764	1.38	0.9162	2	0.9772
0.16	0.5636	0.78	0.7823	1.4	0.9192	2.05	0.9798
0.18	0.5714	0.8	0.7881	1.42	0.9222	2.1	0.9821
0.2	0.5793	0.82	0.7939	1.44	0.9251	2.15	0.9842
0.22	0.5871	0.84	0.7995	1.46	0.9279	2.2	0.9861
0.24	0.5948	0.86	0.8051	1.48	0.9306	2.25	0.9878
0.26	0.6026	0.88	0.8106	1.5	0.9332	2.3	0.9893
0.28	0.6103	0.9	0.8159	1.52	0.9357	2.35	0.9906
0.3	0.6179	0.92	0.8212	1.54	0.9382	2.4	0.9918
0.32	0.6255	0.94	0.8264	1.56	0.9406	2.45	0.9929
0.34	0.6331	0.96	0.8315	1.58	0.9429	2.5	0.9938
0.36	0.6406	0.98	0.8365	1.6	0.9452	2.55	0.9946
0.38	0.6480	1	0.8413	1.62	0.9474	2.6	0.9953
0.4	0.6554	1.02	0.8461	1.64	0.9495	2.65	0.9960
0.42	0.6628	1.04	0.8508	1.66	0.9515	2.7	0.9965
0.44	0.6700	1.06	0.8554	1.68	0.9535	2.75	0.9970
0.46	0.6772	1.08	0.8599	1.7	0.9554	2.8	0.9974
0.48	0.6844	1.1	0.8643	1.72	0.9573	2.85	0.9978
0.5	0.6915	1.12	0.8686	1.74	0.9591	2.9	0.9981
0.52	0.6985	1.14	0.8729	1.76	0.9608	2.95	0.9984
0.54	0.7054	1.16	0.8770	1.78	0.9625	3	0.9987
0.56	0.7123	1.18	0.8810	1.8	0.9641	3.5	0.9998
0.58	0.7190	1.2	0.8849	1.82	0.9656	4	1.0000
0.6	0.7257	1.22	0.8888	1.84	0.9671		

附录四　t 分布临界值表(双侧)

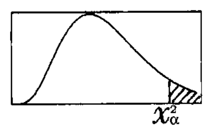
$P\{|t| > t_{\frac{\alpha}{2}}\} = \alpha$



$df \backslash \alpha$	0.001	0.005	0.01	0.02	0.05	0.1	0.2	0.5	0.9
1	636.58	127.32	63.656	31.821	12.706	6.3137	3.0777	1.0000	0.1584
2	31.600	14.089	9.9250	6.9645	4.3027	2.9200	1.8856	0.8165	0.1421
3	12.924	7.4532	5.8408	4.5407	3.1824	2.3534	1.6377	0.7649	0.1366
4	8.6101	5.5975	4.6041	3.7469	2.7765	2.1318	1.5332	0.7407	0.1338
5	6.8685	4.7733	4.0321	3.3649	2.5706	2.0150	1.4759	0.7267	0.1322
6	5.9587	4.3168	3.7074	3.1427	2.4469	1.9432	1.4398	0.7176	0.1311
7	5.4081	4.0294	3.4995	2.9979	2.3646	1.8946	1.4149	0.7111	0.1303
8	5.0414	3.8325	3.3554	2.8965	2.3060	1.8595	1.3968	0.7064	0.1297
9	4.7809	3.6896	3.2498	2.8214	2.2622	1.8331	1.3830	0.7027	0.1293
10	4.5868	3.5814	3.1693	2.7638	2.2281	1.8125	1.3722	0.6998	0.1289
12	4.3178	3.4284	3.0545	2.6810	2.1788	1.7823	1.3562	0.6955	0.1283
14	4.1403	3.3257	2.9768	2.6245	2.1448	1.7613	1.3450	0.6924	0.1280
16	4.0149	3.2520	2.9208	2.5835	2.1199	1.7459	1.3368	0.6901	0.1277
18	3.9217	3.1966	2.8784	2.5524	2.1009	1.7341	1.3304	0.6884	0.1274
20	3.8496	3.1534	2.8453	2.5280	2.0860	1.7247	1.3253	0.6870	0.1273
22	3.7922	3.1188	2.8188	2.5083	2.0739	1.7171	1.3212	0.6858	0.1271
24	3.7454	3.0905	2.7970	2.4922	2.0639	1.7109	1.3178	0.6848	0.1270
26	3.7067	3.0669	2.7787	2.4786	2.0555	1.7056	1.3150	0.6840	0.1269
28	3.6739	3.0470	2.7633	2.4671	2.0484	1.7011	1.3125	0.6834	0.1268
30	3.6460	3.0298	2.7500	2.4573	2.0423	1.6973	1.3104	0.6828	0.1267
35	3.5911	2.9961	2.7238	2.4377	2.0301	1.6896	1.3062	0.6816	0.1266
40	3.5510	2.9712	2.7045	2.4233	2.0211	1.6839	1.3031	0.6807	0.1265
50	3.4960	2.9370	2.6778	2.4033	2.0086	1.6759	1.2987	0.6794	0.1263
60	3.4602	2.9146	2.6603	2.3901	2.0003	1.6706	1.2958	0.6786	0.1262
80	3.4164	2.8870	2.6387	2.3739	1.9901	1.6641	1.2922	0.6776	0.1261
100	3.3905	2.8707	2.6259	2.3642	1.9840	1.6602	1.2901	0.6770	0.1260
∞	3.2906	2.8071	2.5759	2.3264	1.9600	1.6449	1.2816	0.6745	0.1257

附录五　χ^2 分布临界值表

$P\{\chi^2 > \chi^2_\alpha\} = \chi$

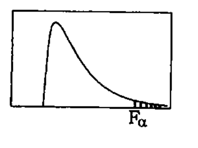


$df \setminus \alpha$	0.99	0.98	0.95	0.90	0.50	0.25	0.10	0.05	0.02	0.01	0.005
1	0.000	0.001	0.004	0.016	0.45	1.32	2.71	3.84	5.41	6.63	7.88
2	0.020	0.040	0.10	0.21	1.39	2.77	4.61	5.99	7.82	9.21	10.60
3	0.11	0.18	0.35	0.58	2.37	4.11	6.25	7.81	9.84	11.34	12.84
4	0.30	0.43	0.71	1.06	3.36	5.39	7.78	9.49	11.67	13.28	14.86
5	0.55	0.75	1.15	1.61	4.35	6.63	9.24	11.07	13.39	15.09	16.75
6	0.87	1.13	1.64	2.20	5.35	7.84	10.64	12.59	15.03	16.81	18.55
7	1.24	1.56	2.17	2.83	6.35	9.04	12.02	14.07	16.62	18.48	20.28
8	1.65	2.03	2.73	3.49	7.34	10.22	13.36	15.51	18.17	20.09	21.95
9	2.09	2.53	3.33	4.17	8.34	11.39	14.68	16.92	19.68	21.67	23.59
10	2.56	3.06	3.94	4.87	9.34	12.55	15.99	18.31	21.16	23.21	25.19
11	3.05	3.61	4.57	5.58	10.34	13.70	17.28	19.68	22.62	24.73	26.76
12	3.57	4.18	5.23	6.30	11.34	14.85	18.55	21.03	24.05	26.22	28.30
13	4.11	4.77	5.89	7.04	12.34	15.98	19.81	22.36	25.47	27.69	29.82
14	4.66	5.37	6.57	7.79	13.34	17.12	21.06	23.68	26.87	29.14	31.32
15	5.23	5.98	7.26	8.55	14.34	18.25	22.31	25.00	28.26	30.58	32.80
16	5.81	6.61	7.96	9.31	15.34	19.37	23.54	26.30	29.63	32.00	34.27
17	6.41	7.25	8.67	10.09	16.34	20.49	24.77	27.59	31.00	33.41	35.72
18	7.01	7.91	9.39	10.86	17.34	21.60	25.99	28.87	32.35	34.81	37.16
19	7.63	8.57	10.12	11.65	18.34	22.72	27.20	30.14	33.69	36.19	38.58
20	8.26	9.24	10.85	12.44	19.34	23.83	28.41	31.41	35.02	37.57	40.00
22	9.54	10.60	12.34	14.04	21.34	26.04	30.81	33.92	37.66	40.29	42.80
24	10.86	11.99	13.85	15.66	23.34	28.24	33.20	36.42	40.27	42.98	45.56
26	12.20	13.41	15.38	17.29	25.34	30.43	35.56	38.89	42.86	45.64	48.29
28	13.56	14.85	16.93	18.94	27.34	32.62	37.92	41.34	45.42	48.28	50.99
30	14.95	16.31	18.49	20.60	29.34	34.80	40.26	43.77	47.96	50.89	53.67
35	18.51	20.03	22.47	24.80	34.34	40.22	46.06	49.80	54.24	57.34	60.27
40	22.16	23.84	26.51	29.05	39.34	45.62	51.81	55.76	60.44	63.69	66.77
50	29.71	31.66	34.76	37.69	49.33	56.33	63.17	67.50	72.61	76.15	79.49
60	37.48	39.70	43.19	46.46	59.33	66.98	74.40	79.08	84.58	88.38	91.95
80	53.54	56.21	60.39	64.28	79.33	88.13	96.58	101.9	108.1	112.3	116.3
100	70.06	73.14	77.93	82.36	99.3	109.1	118.5	124.3	131.1	135.8	140.2

附录六　F 分布临界值表

$$P(F > F_{\alpha}) = \alpha$$

$$\alpha = 0.1$$



df1/df2	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30	40	60	100	200	400	∞
1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.1	61.3	61.6	61.7	62.3	62.5	62.8	63.0	63.2	63.2	63.3
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.43	9.44	9.44	9.46	9.47	9.47	9.48	9.49	9.49	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.20	5.19	5.18	5.17	5.16	5.15	5.14	5.14	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.88	3.86	3.85	3.84	3.82	3.80	3.79	3.78	3.77	3.77	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.25	3.23	3.22	3.21	3.17	3.16	3.14	3.13	3.12	3.11	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.88	2.86	2.85	2.84	2.80	2.78	2.76	2.75	2.73	2.73	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.64	2.62	2.61	2.59	2.56	2.54	2.51	2.50	2.48	2.48	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.48	2.45	2.44	2.42	2.38	2.36	2.34	2.32	2.31	2.30	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.35	2.33	2.31	2.30	2.25	2.23	2.21	2.19	2.17	2.17	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.26	2.23	2.22	2.20	2.16	2.13	2.11	2.09	2.07	2.06	2.06
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.12	2.09	2.08	2.06	2.01	1.99	1.96	1.94	1.92	1.91	1.90
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.02	2.00	1.98	1.96	1.91	1.89	1.86	1.83	1.82	1.81	1.80
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.95	1.93	1.91	1.89	1.84	1.81	1.78	1.76	1.74	1.73	1.72
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.90	1.87	1.85	1.84	1.78	1.75	1.72	1.70	1.68	1.67	1.66
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.86	1.83	1.81	1.79	1.74	1.71	1.68	1.65	1.63	1.62	1.61
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.74	1.71	1.69	1.67	1.61	1.57	1.54	1.51	1.48	1.47	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.68	1.65	1.62	1.61	1.54	1.51	1.47	1.43	1.41	1.39	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.62	1.59	1.56	1.54	1.48	1.44	1.40	1.36	1.33	1.31	1.29
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.61	1.57	1.54	1.52	1.49	1.42	1.38	1.34	1.29	1.26	1.24	1.21
200	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.58	1.54	1.51	1.48	1.46	1.38	1.34	1.29	1.24	1.20	1.17	1.14
400	2.72	2.32	2.10	1.96	1.86	1.79	1.73	1.69	1.65	1.61	1.56	1.52	1.49	1.46	1.44	1.36	1.32	1.26	1.21	1.17	1.14	1.10
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.50	1.47	1.44	1.42	1.34	1.30	1.24	1.18	1.13	1.09	1.00

F 分布临界值表(续一)

$\alpha = 0.05$

df1/ df2	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30	40	60	100	200	400	∞
1	161	199	216	225	230	234	237	239	241	242	244	245	246	247	248	250	251	252	253	254	254	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66	8.62	8.59	8.57	8.55	8.54	8.53	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80	5.75	5.72	5.69	5.66	5.65	5.64	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56	4.50	4.46	4.43	4.41	4.39	4.38	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87	3.81	3.77	3.74	3.71	3.69	3.68	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44	3.38	3.34	3.30	3.27	3.25	3.24	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15	3.08	3.04	3.01	2.97	2.95	2.94	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94	2.86	2.83	2.79	2.76	2.73	2.72	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77	2.70	2.66	2.62	2.59	2.56	2.55	2.54
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54	2.47	2.43	2.38	2.35	2.32	2.31	2.30
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39	2.31	2.27	2.22	2.19	2.16	2.15	2.13
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28	2.19	2.15	2.11	2.07	2.04	2.02	2.01
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19	2.11	2.06	2.02	1.98	1.95	1.93	1.92
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.22	2.18	2.15	2.12	2.04	1.99	1.95	1.91	1.88	1.86	1.84
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93	1.84	1.79	1.74	1.70	1.66	1.64	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84	1.74	1.69	1.64	1.59	1.55	1.53	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75	1.65	1.59	1.53	1.48	1.44	1.41	1.39
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68	1.57	1.52	1.45	1.39	1.34	1.31	1.28
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.66	1.62	1.52	1.46	1.39	1.32	1.26	1.23	1.19
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.78	1.72	1.67	1.63	1.60	1.49	1.42	1.35	1.28	1.22	1.18	1.13
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.69	1.64	1.60	1.57	1.46	1.39	1.32	1.24	1.17	1.12	1.00

F 分布临界值表(续二)

$\alpha = 0.02$

df1/df2	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30	40	60	100	200	400	∞
1	1013	1249	1351	1406	1441	1464	1482	1495	1505	1514	1526	1535	1542	1548	1552	1565	1571	1578	1583	1587	1589	1591
2	48.5	49.0	49.2	49.2	49.3	49.3	49.4	49.4	49.4	49.4	49.4	49.4	49.4	49.4	49.4	49.5	49.5	49.5	49.5	49.5	49.5	49.5
3	20.6	18.9	18.1	17.7	17.4	17.2	17.1	17.0	16.9	16.9	16.8	16.7	16.6	16.6	16.6	16.4	16.4	16.3	16.3	16.3	16.2	16.2
4	14.0	12.1	11.3	10.9	10.6	10.4	10.3	10.2	10.1	10.0	9.89	9.81	9.75	9.71	9.67	9.55	9.50	9.44	9.39	9.35	9.33	9.32
5	11.3	9.45	8.67	8.23	7.95	7.76	7.61	7.50	7.42	7.34	7.23	7.16	7.10	7.05	7.01	6.89	6.83	6.77	6.72	6.69	6.67	6.65
6	9.88	8.05	7.29	6.86	6.58	6.39	6.25	6.14	6.05	5.98	5.88	5.80	5.74	5.69	5.65	5.53	5.47	5.41	5.36	5.33	5.31	5.29
7	8.99	7.20	6.45	6.03	5.76	5.58	5.44	5.33	5.24	5.17	5.06	4.98	4.92	4.88	4.84	4.72	4.66	4.60	4.55	4.51	4.49	4.47
8	8.39	6.64	5.90	5.49	5.22	5.04	4.90	4.79	4.70	4.63	4.53	4.45	4.39	4.34	4.30	4.19	4.13	4.06	4.01	3.97	3.96	3.94
9	7.96	6.23	5.51	5.10	4.84	4.65	4.52	4.41	4.33	4.26	4.15	4.07	4.01	3.96	3.92	3.81	3.75	3.68	3.63	3.59	3.57	3.55
10	7.64	5.93	5.22	4.82	4.55	4.37	4.23	4.13	4.04	3.97	3.87	3.79	3.73	3.68	3.64	3.52	3.46	3.40	3.35	3.31	3.29	3.27
12	7.19	5.52	4.81	4.42	4.16	3.98	3.85	3.74	3.66	3.59	3.48	3.40	3.34	3.29	3.25	3.13	3.07	3.01	2.95	2.91	2.89	2.87
14	6.89	5.24	4.55	4.16	3.90	3.72	3.59	3.48	3.40	3.33	3.23	3.15	3.09	3.04	3.00	2.88	2.81	2.75	2.69	2.65	2.63	2.61
16	6.67	5.05	4.36	3.97	3.72	3.54	3.41	3.30	3.22	3.15	3.05	2.97	2.90	2.86	2.82	2.69	2.63	2.56	2.51	2.46	2.44	2.42
18	6.51	4.90	4.22	3.84	3.59	3.41	3.27	3.17	3.09	3.02	2.91	2.83	2.77	2.72	2.68	2.56	2.49	2.42	2.37	2.32	2.30	2.28
20	6.39	4.79	4.11	3.73	3.48	3.30	3.17	3.07	2.98	2.91	2.81	2.73	2.67	2.62	2.58	2.45	2.38	2.31	2.26	2.21	2.19	2.17
25	6.18	4.59	3.93	3.55	3.30	3.13	2.99	2.89	2.81	2.74	2.63	2.55	2.49	2.44	2.40	2.27	2.20	2.13	2.07	2.02	1.99	1.97
30	6.04	4.47	3.81	3.43	3.19	3.01	2.88	2.78	2.69	2.62	2.52	2.44	2.37	2.32	2.28	2.15	2.08	2.00	1.94	1.89	1.87	1.84
40	5.87	4.32	3.67	3.30	3.05	2.88	2.74	2.64	2.56	2.49	2.38	2.30	2.23	2.18	2.14	2.01	1.93	1.86	1.79	1.74	1.71	1.68
60	5.71	4.18	3.53	3.16	2.92	2.75	2.62	2.51	2.43	2.36	2.25	2.17	2.10	2.05	2.01	1.87	1.79	1.71	1.64	1.58	1.54	1.51
100	5.59	4.07	3.43	3.06	2.82	2.65	2.52	2.41	2.33	2.26	2.15	2.07	2.00	1.95	1.90	1.76	1.68	1.59	1.51	1.44	1.41	1.37
200	5.50	3.99	3.35	2.99	2.75	2.58	2.44	2.34	2.26	2.19	2.08	1.99	1.93	1.87	1.83	1.68	1.60	1.50	1.41	1.34	1.29	1.24
400	5.46	3.95	3.32	2.95	2.71	2.54	2.41	2.31	2.22	2.15	2.04	1.96	1.89	1.83	1.79	1.64	1.55	1.46	1.36	1.28	1.23	1.16
∞	5.41	3.91	3.28	2.92	2.68	2.51	2.37	2.27	2.19	2.12	2.00	1.92	1.85	1.80	1.75	1.60	1.51	1.41	1.31	1.22	1.15	1.00

F分布临界值表(续三)

$\alpha = 0.01$

df1/df2	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30	40	60	100	200	400	∞
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6143	6170	6191	6209	6260	6286	6313	6334	6350	6358	6366
2.0	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5
3.0	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.8	26.8	26.7	26.5	26.4	26.3	26.2	26.2	26.2	26.1
4.0	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.2	14.1	14.0	13.8	13.7	13.7	13.6	13.5	13.5	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.77	9.68	9.61	9.55	9.38	9.29	9.20	9.13	9.08	9.05	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.60	7.52	7.45	7.40	7.23	7.14	7.06	6.99	6.93	6.91	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.36	6.28	6.21	6.16	5.99	5.91	5.82	5.75	5.70	5.68	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.56	5.48	5.41	5.36	5.20	5.12	5.03	4.96	4.91	4.89	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	5.01	4.92	4.86	4.81	4.65	4.57	4.48	4.41	4.36	4.34	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.60	4.52	4.46	4.41	4.25	4.17	4.08	4.01	3.96	3.94	3.91
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.05	3.97	3.91	3.86	3.70	3.62	3.54	3.47	3.41	3.39	3.36
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.70	3.62	3.56	3.51	3.35	3.27	3.18	3.11	3.06	3.03	3.00
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.45	3.37	3.31	3.26	3.10	3.02	2.93	2.86	2.81	2.78	2.75
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.27	3.19	3.13	3.08	2.92	2.84	2.75	2.68	2.62	2.59	2.57
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.13	3.05	2.99	2.94	2.78	2.69	2.61	2.54	2.48	2.45	2.42
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.74	2.66	2.60	2.55	2.39	2.30	2.21	2.13	2.07	2.04	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.56	2.48	2.42	2.37	2.20	2.11	2.02	1.94	1.87	1.84	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.39	2.31	2.25	2.20	2.03	1.94	1.84	1.75	1.68	1.64	1.60
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.27	2.19	2.12	2.07	1.89	1.80	1.69	1.60	1.52	1.47	1.43
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.17	2.09	2.03	1.97	1.79	1.69	1.58	1.48	1.39	1.34	1.28
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.23	2.13	2.05	1.98	1.92	1.75	1.64	1.53	1.42	1.32	1.26	1.19
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.08	2.00	1.93	1.88	1.70	1.59	1.47	1.36	1.25	1.17	1.00

F 分布临界值表(续四)

$\alpha = 0.001$ (“*”表示乘 100)

df1/ df2	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30	40	60	100	200	400	∞
1	4053*	4997*	5402*	5626*	5764*	5860*	5931*	5979*	6022*	6055*	6103*	6141*	6170*	6189*	6208*	6260*	6284*	6313*	6332*	6351*	6361*	6365*
2	998	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
3	167	148	141	137	135	133	132	131	130	129	128	128	127	127	126	125	125	124	124	124	124	123
4	74.1	61.2	56.2	53.4	51.7	50.5	49.7	49.0	48.5	48.1	47.4	46.9	46.6	46.3	46.1	45.4	45.1	44.7	44.5	44.3	44.2	44.0
5	47.2	37.1	33.2	31.1	29.8	28.8	28.2	27.6	27.2	26.9	26.4	26.1	25.8	25.6	25.4	24.9	24.6	24.3	24.1	24.0	23.9	23.8
6	35.5	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.7	18.4	18.0	17.7	17.5	17.3	17.1	16.7	16.4	16.2	16.0	15.9	15.8	15.7
7	29.2	21.7	18.8	17.2	16.2	15.5	15.0	14.6	14.3	14.1	13.7	13.4	13.2	13.1	12.9	12.5	12.3	12.1	12.0	11.8	11.8	11.7
8	25.4	18.5	15.8	14.4	13.5	12.9	12.4	12.0	11.8	11.5	11.2	10.9	10.8	10.6	10.5	10.1	9.92	9.73	9.57	9.45	9.39	9.33
9	22.9	16.4	13.9	12.6	11.7	11.1	10.7	10.4	10.1	9.89	9.57	9.33	9.15	9.01	8.90	8.55	8.37	8.19	8.04	7.93	7.87	7.81
10	21.0	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.96	8.75	8.45	8.22	8.05	7.91	7.80	7.47	7.30	7.12	6.98	6.87	6.82	6.76
12	18.6	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.79	6.63	6.51	6.40	6.09	5.93	5.76	5.63	5.52	5.47	5.42
14	17.1	11.8	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	6.13	5.93	5.78	5.66	5.56	5.25	5.10	4.94	4.81	4.71	4.66	4.60
16	16.1	11.0	9.01	7.94	7.27	6.80	6.46	6.20	5.98	5.81	5.55	5.35	5.21	5.09	4.99	4.70	4.54	4.39	4.26	4.16	4.11	4.06
18	15.4	10.4	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	5.13	4.94	4.80	4.68	4.59	4.30	4.15	4.00	3.87	3.77	3.72	3.67
20	14.8	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.64	4.49	4.38	4.29	4.00	3.86	3.70	3.58	3.48	3.43	3.38
30	13.3	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.82	3.69	3.58	3.49	3.22	3.07	2.92	2.79	2.69	2.64	2.59
40	12.6	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.64	3.47	3.34	3.23	3.15	2.87	2.73	2.57	2.44	2.34	2.29	2.23
60	12.0	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54	3.32	3.15	3.02	2.91	2.83	2.55	2.41	2.25	2.12	2.01	1.95	1.89
100	11.5	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	3.07	2.91	2.78	2.68	2.59	2.32	2.17	2.01	1.87	1.75	1.68	1.62
200	11.2	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26	3.12	2.90	2.74	2.61	2.51	2.42	2.15	2.00	1.83	1.68	1.55	1.48	1.39
400	11.0	7.03	5.53	4.71	4.19	3.83	3.56	3.35	3.18	3.04	2.82	2.66	2.53	2.43	2.34	2.07	1.92	1.75	1.59	1.45	1.36	1.26
∞	10.8	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96	2.74	2.58	2.45	2.35	2.27	1.99	1.84	1.66	1.49	1.34	1.23	1.00

索　引

（以汉语拼音排序）

S

T

X

Y

Z